Computational Systems Biology
Deep Learning in the Life Sciences

6.802  20.390  20.490  HST.506
6.874 Area II TQE (AI)

David Gifford
Lecture 1
February 4, 2019

http://mit6874.github.io
mit6874.github.io
6.874staff@mit.edu

Please use Piazza or the staff email for any questions

You should have received the Google Cloud coupon URL in your email
Teaching Staff

David Gifford
Gifford@mit.edu

Manolis Kellis
manoli@mit.edu

Tim Truong
ttruong@mit.edu

Sachit Saksena
sachit@mit.edu

Corban Swain
c_swain@mit.edu
Recitations (this week)
Thursday 4 - 5pm 36-156
Friday 4 - 5pm 36-156

Office hours are after recitation at 5pm in same room
(PS1 help and advice)
Approximately 8% of deep learning publications are in bioinformatics.
Welcome to a new approach to life sciences research

• Enabled by the convergence of three things
  • Inexpensive, high-quality, collection of large data sets (sequencing, imaging, etc.)
  • New machine learning methods (including ensemble methods)
  • High-performance Graphics Processing Unit (GPU) machine learning implementations
• Result is completely transformative
Your background

- Calculus, Linear Algebra
- Probability, Programming
- Introductory Biology
Grade contributions

- Four Problem Sets (40%)
  - Individual contribution
  - Done using Google Cloud, Jupyter Notebook
- Two quizzes (1.5 hours), one sheet of notes (30%)
- Final Project (25%)
  - Done in teams of two
- Scribing (5%)
Alternative MIT subjects

- 6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
- 6.S897/HST.956: Machine Learning for Healthcare (2:30pm 4-270)
- 8.592 Statistical Physics in Biology
- 7.09 Quantitative and Computational Biology
- 7.32 Systems Biology
- 7.33 Evolutionary Biology: Concepts, Models and Computation
- 7.57 Quantitative Biology for Graduate Students
- 18.417 Introduction to Computational Molecular Biology
- 20.482 Foundations of Algorithms and Computational Techniques in Systems Biology
<table>
<thead>
<tr>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Module</th>
<th>Lec/Rec</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS1: Softmax warmup (MNIST) (out: Tue 2/6, due: Fri 2/21)</td>
<td>1</td>
<td>Tuesday, February 4, 2020</td>
<td>Module 1: ML models and interpretation</td>
<td>Lecture 1</td>
<td>Scope of the subject, ML Intro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 6, 2020</td>
<td></td>
<td>Lecture 2</td>
<td>Learning MLPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 7, 2020</td>
<td></td>
<td>Recitation 1</td>
<td>ML and Google notebook overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, February 11, 2020</td>
<td>Module 2: Chromatin structure / Model selection and uncertainty</td>
<td>Lecture 3</td>
<td>Model capacity hypothesis space, Neural Networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 13, 2020</td>
<td></td>
<td>Lecture 4</td>
<td>Convolutional neural networks, Recurrent neural networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 14, 2020</td>
<td></td>
<td>Recitation 2</td>
<td>Neural Networks Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, February 18, 2020</td>
<td></td>
<td></td>
<td>(Holiday - President’s Day)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 20, 2020</td>
<td>Lecture 5: ML model interpretation I (SIS) (Brandon Carter Guest Lecture)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 21, 2020</td>
<td>Recitation 3: Interpreting ML models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PS2: TF Binding, ChIP, Motifs (out: Fri 2/21, Due: Fri 3/13)</td>
<td>4</td>
<td>Tuesday, February 25, 2020</td>
<td>Module 3: Expressed Genome / Dimensionality reduction</td>
<td>Lecture 6</td>
<td>Chromatin accessibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 27, 2020</td>
<td></td>
<td>Lecture 7</td>
<td>Protein-DNA interactions and ChIP seq motif discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 28, 2020</td>
<td></td>
<td>Lecture 8</td>
<td>Model uncertainty and experiment design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, March 3, 2020</td>
<td></td>
<td>Lecture 9</td>
<td>Generative models (gradients, VAEs, GANs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 5, 2020</td>
<td></td>
<td>Recitation 4</td>
<td>Chromatin and gene regulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 6, 2020</td>
<td></td>
<td>Recitation 5</td>
<td>Model uncertainty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, March 10, 2020</td>
<td></td>
<td>Lecture 10</td>
<td>Chromatin interactions and 3D genome organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 12, 2020</td>
<td></td>
<td>Lecture 11</td>
<td>Dimensionality reduction (PCA, t-SNE, autoencoders)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 13, 2020</td>
<td></td>
<td>Recitation 6</td>
<td>Regulatory element models</td>
</tr>
<tr>
<td>PS3: scRNA-seq tSNE analysis (out: Thu 3/12, due Fri 4/3)</td>
<td>7</td>
<td>Tuesday, March 17, 2020</td>
<td>Module 4: Human Genetics - Genotype -&gt; Phenotype</td>
<td>Lecture 12</td>
<td>The expressed genome and RNA splicing (RNA-seq)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 19, 2020</td>
<td></td>
<td>Lecture 13</td>
<td>Quiz 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 20, 2020</td>
<td></td>
<td>Recitation 7</td>
<td>No recitation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, March 31, 2020</td>
<td></td>
<td>Lecture 14</td>
<td>scRNA-seq and cell labeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 2, 2020</td>
<td></td>
<td>Lecture 15</td>
<td>Manifolds, manifold mapping, word2vec</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 3, 2020</td>
<td></td>
<td>Recitation 8</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td>PS4: Disease, genetics, diagnostics (Out: Fri 4/3, Due: Fri 4/17)</td>
<td>10</td>
<td>Tuesday, April 7, 2020</td>
<td>Module 5: Therapeutics and Diagnostics</td>
<td>Lecture 16</td>
<td>Deep learning in Disease Studies and Human Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 9, 2020</td>
<td></td>
<td>Lecture 17</td>
<td>eQTL prediction and variant prioritization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 10, 2020</td>
<td></td>
<td>Recitation 9</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, April 14, 2020</td>
<td></td>
<td>Lecture 18</td>
<td>STARR-seq and GWAS studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 16, 2020</td>
<td></td>
<td>Lecture 19</td>
<td>High-throughput experimentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 17, 2020</td>
<td></td>
<td>Recitation 10</td>
<td>Protein structure prediction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, April 21, 2020</td>
<td>Projects</td>
<td>Lecture 20</td>
<td>Therapeutic design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 23, 2020</td>
<td></td>
<td>Lecture 21</td>
<td>Imaging and genotype to phenotype (Guest: Adrian Dalca)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 24, 2020</td>
<td></td>
<td>Recitation 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, April 28, 2020</td>
<td></td>
<td>Lecture 22</td>
<td>Quiz 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 30, 2020</td>
<td></td>
<td>Lecture 23</td>
<td>How to write, how to present</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, May 5, 2020</td>
<td>Projects</td>
<td>Recitation 12</td>
<td>(Project work)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, May 7, 2020</td>
<td></td>
<td>Lecture 24</td>
<td>Project Presentations I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, May 12, 2020</td>
<td>Projects</td>
<td>Lecture 25</td>
<td>Project Presentations II</td>
</tr>
</tbody>
</table>

Modules:
- **Module 1**: ML models and interpretation
- **Module 2**: Chromatin structure / Model selection and uncertainty
- **Module 3**: Expressed Genome / Dimensionality reduction
- **Module 4**: Human Genetics - Genotype -> Phenotype
- **Module 5**: Therapeutics and Diagnostics

Projects:
- PS1: Softmax warmup (MNIST) (out: Tue 2/6, due: Fri 2/21)
- PS2: TF Binding, ChIP, Motifs (out: Fri 2/21, Due: Fri 3/13)
- PS3: scRNA-seq tSNE analysis (out: Thu 3/12, due Fri 4/3)
- PS4: Disease, genetics, diagnostics (Out: Fri 4/3, Due: Fri 4/17)

Lectures and Recitations:
- Lecture 1: Scope of the subject, ML Intro
- Lecture 2: Learning MLPs
- Recitation 1: ML and Google notebook overview
- Lecture 3: Model capacity hypothesis space, Neural Networks
- Lecture 4: Convolutional neural networks, Recurrent neural networks
- Recitation 2: Neural Networks Review
- Lecture 5: ML model interpretation I (SIS) (Brandon Carter Guest Lecture)
- Recitation 6: Interpreting ML models
- Lecture 6: Chromatin accessibility
- Lecture 7: Protein-DNA interactions and ChIP seq motif discovery
- Lecture 8: Model uncertainty and experiment design
- Lecture 9: Generative models (gradients, VAEs, GANs)
- Lecture 10: Chromatin interactions and 3D genome organization
- Lecture 11: Dimensionality reduction (PCA, t-SNE, autoencoders)
- Lecture 12: The expressed genome and RNA splicing (RNA-seq)
- Lecture 13: Quiz 1
- Lecture 14: scRNA-seq and cell labeling
- Lecture 15: Manifolds, manifold mapping, word2vec
- Lecture 16: Deep learning in Disease Studies and Human Genetics
- Lecture 17: eQTL prediction and variant prioritization
- Lecture 18: STARR-seq and GWAS studies
- Lecture 19: High-throughput experimentation
- Lecture 20: Therapeutic design
- Lecture 21: Imaging and genotype to phenotype (Guest: Adrian Dalca)
- Lecture 22: Quiz 2
- Lecture 23: How to write, how to present
- Lecture 24: Project Presentations I
- Lecture 25: Project Presentations II
PS 1: TensorFlow Warm Up

ground truth: 5
<table>
<thead>
<tr>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Module</th>
<th>Lec/Rec</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS1: Softmax warmup (MNIST)</td>
<td>1</td>
<td>Tuesday, February 4, 2020</td>
<td>Module 1: ML models and interpretation</td>
<td>Lecture 1</td>
<td>Scope of the subject, ML Intro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 7, 2020</td>
<td></td>
<td>Lecture 2</td>
<td>Learning MLPs</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Tuesday, February 11, 2020</td>
<td></td>
<td>Recitation 1</td>
<td>ML and Google notebook overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 13, 2020</td>
<td></td>
<td>Lecture 3</td>
<td>Model capacity hypothesis space, Neural Networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 14, 2020</td>
<td></td>
<td>Lecture 4</td>
<td>Convolutional neural networks, Recurrent neural networks</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Tuesday, February 18, 2020</td>
<td></td>
<td>Recitation 2</td>
<td>Neural Networks Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 20, 2020</td>
<td></td>
<td>Lecture 5</td>
<td>ML model interpretation I (SIS) (Brandon Carter Guest Lecture)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 21, 2020</td>
<td></td>
<td>Recitation 3</td>
<td>Interpreting ML models</td>
</tr>
<tr>
<td>PS2: TF Binding, ChIP, Motifs</td>
<td>4</td>
<td>Tuesday, February 25, 2020</td>
<td>Module 2: Chromatin structure / Model selection and uncertainty</td>
<td>Lecture 6</td>
<td>Chromatin accessibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 27, 2020</td>
<td></td>
<td>Lecture 7</td>
<td>Protein-DNA interactions and ChIP seq motif discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 28, 2020</td>
<td></td>
<td>Recitation 4</td>
<td>Chromatin and gene regulation</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Tuesday, March 5, 2020</td>
<td></td>
<td>Lecture 8</td>
<td>Model uncertainty and experiment design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 6, 2020</td>
<td></td>
<td>Lecture 9</td>
<td>Generative models (gradients, VAEs, GANs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuesday, March 10, 2020</td>
<td></td>
<td>Recitation 5</td>
<td>Model uncertainty</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Thursday, March 12, 2020</td>
<td></td>
<td>Lecture 10</td>
<td>Chromatin interactions and 3D genome organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 13, 2020</td>
<td></td>
<td>Lecture 11</td>
<td>Dimensionality reduction (PCA, t-SNE, autoencoders)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 17, 2020</td>
<td></td>
<td>Lecture 12</td>
<td>The expressed genome and RNA splicing (RNA-seq)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Friday, March 20, 2020</td>
<td></td>
<td>Lecture 13</td>
<td>Quiz 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 19, 2020</td>
<td></td>
<td>Recitation 7</td>
<td>No recitation</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Tuesday, March 24, 2020</td>
<td>Module 3: Expressed Genome / Dimensionality reduction</td>
<td>Lecture 14</td>
<td>scRNA-seq and cell labeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 26, 2020</td>
<td></td>
<td>Lecture 15</td>
<td>Manifolds, manifold mapping, word2vec</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 20, 2020</td>
<td></td>
<td>Recitation 8</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Tuesday, March 31, 2020</td>
<td></td>
<td>Lecture 16</td>
<td>Deep learning in Disease Studies and Human Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 2, 2020</td>
<td></td>
<td>Lecture 17</td>
<td>eQTL prediction and variant prioritization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 3, 2020</td>
<td></td>
<td>Recitation 9</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Tuesday, April 7, 2020</td>
<td>Module 4: Human Genetics - Genotype -&gt; Phenotype</td>
<td>Lecture 18</td>
<td>STARR-seq and GWAS studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 9, 2020</td>
<td></td>
<td>Lecture 19</td>
<td>High-throughput experimentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 10, 2020</td>
<td></td>
<td>Recitation 10</td>
<td>Protein structure prediction</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Tuesday, April 14, 2020</td>
<td></td>
<td>Lecture 20</td>
<td>Therapeutic design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 16, 2020</td>
<td></td>
<td>Lecture 21</td>
<td>Imaging and genotype to phenotype (Guest: Adrian Dalca)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 17, 2020</td>
<td></td>
<td>Recitation 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Tuesday, April 21, 2020</td>
<td>Module 5: Therapeutics and Diagnostics</td>
<td>Lecture 22</td>
<td>Quiz 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 23, 2020</td>
<td></td>
<td>Lecture 23</td>
<td>How to write, how to present</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 24, 2020</td>
<td></td>
<td>Lecture 24</td>
<td>Project Presentations I</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Tuesday, April 28, 2020</td>
<td>Projects</td>
<td>Lecture 25</td>
<td>Project Presentations II</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, May 5, 2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Tuesday, May 7, 2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Tuesday, May 12, 2020</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PS 2: Genomic regulatory codes
<table>
<thead>
<tr>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Module</th>
<th>Lec/Rec</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>Tuesday, February 4, 2020</td>
<td>Lecture 1</td>
<td>Scope of the subject, ML Intro</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, February 6, 2020</td>
<td>Lecture 2</td>
<td>Learning MLPs</td>
</tr>
<tr>
<td>PS1: Softmax warmup (MNIST) (out: Tue 2/6, due: Fri 2/21)</td>
<td></td>
<td></td>
<td>Friday, February 7, 2020</td>
<td>Recitation 1</td>
<td>ML and Google notebook overview</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>Tuesday, February 11, 2020</td>
<td>Lecture 3</td>
<td>Model capacity hypothesis space, Neural Networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, February 13, 2020</td>
<td>Lecture 4</td>
<td>Convolutional neural networks, Recurrent neural networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, February 14, 2020</td>
<td>Recitation 2</td>
<td>Neural Networks Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tuesday, February 18, 2020</td>
<td></td>
<td>(Holiday - President’s Day)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>Thursday, February 20, 2020</td>
<td>Lecture 5</td>
<td>ML model interpretation I (SIS) (Brandon Carter Guest Lecture)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, February 21, 2020</td>
<td>Recitation 3</td>
<td>Interpreting ML models</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>Tuesday, February 25, 2020</td>
<td>Lecture 6</td>
<td>Chromatin accessibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, February 27, 2020</td>
<td>Lecture 7</td>
<td>Protein-DNA interactions and ChIP seq motif discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, February 28, 2020</td>
<td>Recitation 4</td>
<td>Chromatin and gene regulation</td>
</tr>
<tr>
<td>PS2: TF Binding, ChIP, Motifs (out: Fri 2/21, Due: Fri 3/13)</td>
<td>5</td>
<td></td>
<td>Tuesday, March 3, 2020</td>
<td>Lecture 8</td>
<td>Model uncertainty and experiment design</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, March 5, 2020</td>
<td>Lecture 9</td>
<td>Generative models (gradients, VAEs, GANs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, March 6, 2020</td>
<td>Recitation 5</td>
<td>Model uncertainty</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>Tuesday, March 10, 2020</td>
<td>Lecture 10</td>
<td>Chromatin interactions and 3D genome organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, March 12, 2020</td>
<td>Lecture 11</td>
<td>Dimensionality reduction (PCA, t-SNE, autoencoders)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, March 13, 2020</td>
<td>Recitation 6</td>
<td>Regulatory element models</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>Tuesday, March 17, 2020</td>
<td>Lecture 12</td>
<td>The expressed genome and RNA splicing (RNA-seq)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, March 19, 2020</td>
<td>Lecture 13</td>
<td>Quiz 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, March 20, 2020</td>
<td>Recitation 7</td>
<td>No recitation</td>
</tr>
<tr>
<td>PS3: scRNA-seq tSNE analysis (out: Thu 3/12, due Fri 4/3)</td>
<td>8</td>
<td></td>
<td>Tuesday, March 24, 2020</td>
<td></td>
<td>(Spring Vacation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, March 26, 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, March 20, 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td>Tuesday, March 31, 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, April 2, 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, April 3, 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td>Tuesday, April 7, 2020</td>
<td>Lecture 14</td>
<td>scRNA-seq and cell labeling</td>
</tr>
<tr>
<td>PS4: Disease, genetics, diagnostics (Out: Fri 4/3, Due: Fri 4/17)</td>
<td></td>
<td></td>
<td>Thursday, April 9, 2020</td>
<td>Lecture 15</td>
<td>Manifolds, manifold mapping, word2vec</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, April 10, 2020</td>
<td>Recitation 8</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>Tuesday, April 14, 2020</td>
<td>Lecture 16</td>
<td>Deep learning in Disease Studies and Human Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, April 16, 2020</td>
<td>Lecture 17</td>
<td>eQTL prediction and variant prioritization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, April 17, 2020</td>
<td>Recitation 9</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lecture 18</td>
<td>STARR-seq and GWAS studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lecture 19</td>
<td>High-throughput experimentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recitation 10</td>
<td>Protein structure prediction</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td>Tuesday, April 21, 2020</td>
<td>Lecture 20</td>
<td>Therapeutic design</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, April 23, 2020</td>
<td>Lecture 21</td>
<td>Imaging and genotype to phenotype (Guest: Adrian Dalca)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Friday, April 24, 2020</td>
<td>Recitation 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td>Tuesday, April 28, 2020</td>
<td>Lecture 22</td>
<td>Quiz 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, April 30, 2020</td>
<td>Lecture 23</td>
<td>How to write, how to present</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recitation 12</td>
<td>(Project work)</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td>Tuesday, May 5, 2020</td>
<td>Lecture 24</td>
<td>Project Presentations I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thursday, May 7, 2020</td>
<td>Lecture 25</td>
<td>Project Presentations II</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td>Tuesday, May 12, 2020</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Modules

**Module 1: ML models and interpretation**
- Lecture 1: Scope of the subject, ML Intro
- Lecture 2: Learning MLPs
- Recitation 1: ML and Google notebook overview
- Lecture 3: Model capacity hypothesis space, Neural Networks
- Lecture 4: Convolutional neural networks, Recurrent neural networks
- Recitation 2: Neural Networks Review
- (Holiday - President’s Day)
- Lecture 5: ML model interpretation I (SIS) (Brandon Carter Guest Lecture)
- Recitation 3: Interpreting ML models

**Module 2: Chromatin structure / Model selection and uncertainty**
- Lecture 6: Chromatin accessibility
- Lecture 7: Protein-DNA interactions and ChIP seq motif discovery
- Recitation 4: Chromatin and gene regulation
- Lecture 8: Model uncertainty and experiment design
- Lecture 9: Generative models (gradients, VAEs, GANs)
- Recitation 5: Model uncertainty
- Lecture 10: Chromatin interactions and 3D genome organization
- Lecture 11: Dimensionality reduction (PCA, t-SNE, autoencoders)
- Recitation 6: Regulatory element models
- Lecture 12: The expressed genome and RNA splicing (RNA-seq)
- Lecture 13: Quiz 1
- Recitation 7: No recitation

**Module 3: Expressed Genome / Dimensionality reduction**
- Lecture 14: scRNA-seq and cell labeling
- Lecture 15: Manifolds, manifold mapping, word2vec
- Recitation 8: Dimensionality reduction
- Lecture 16: Deep learning in Disease Studies and Human Genetics
- Lecture 17: eQTL prediction and variant prioritization
- Recitation 9: Genetics
- Lecture 18: STARR-seq and GWAS studies
- Lecture 19: High-throughput experimentation
- Recitation 10: Protein structure prediction

**Module 4: Human Genetics - Genotype -> Phenotype**
- Lecture 20: Therapeutic design
- Lecture 21: Imaging and genotype to phenotype (Guest: Adrian Dalca)
- Recitation 11:

**Module 5: Therapeutics and Diagnostics**
- Lecture 22: Quiz 2
- Lecture 23: How to write, how to present
- Lecture 24: Project Presentations I
- Lecture 25: Project Presentations II

**Projects**
- Lecture 26: Project work

**Psets**
- PS1: Softmax warmup (MNIST) (out: Tue 2/6, due: Fri 2/21)
- PS2: TF Binding, ChIP, Motifs (out: Fri 2/21, Due: Fri 3/13)
- PS3: scRNA-seq tSNE analysis (out: Thu 3/12, due Fri 4/3)
- PS4: Disease, genetics, diagnostics (Out: Fri 4/3, Due: Fri 4/17)
- No other psets
PS 3: Parametric tSNE
Single Cell RNA-seq data
<table>
<thead>
<tr>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Module</th>
<th>Lec/Rec</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS1: Softmax warmup (MNIST) (out: Tue 2/6, due: Fri 2/21)</td>
<td>1</td>
<td>Tuesday, February 4, 2020</td>
<td></td>
<td>Lecture 1</td>
<td>Scope of the subject, ML Intro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 6, 2020</td>
<td></td>
<td>Lecture 2</td>
<td>Learning MLPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 7, 2020</td>
<td></td>
<td>Recitation 1</td>
<td>ML and Google notebook overview</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Tuesday, February 11, 2020</td>
<td>Module 1: ML models and interpretation</td>
<td>Lecture 3</td>
<td>Model capacity hypothesis space, Neural Networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 13, 2020</td>
<td></td>
<td>Lecture 4</td>
<td>Convolutional neural networks, Recurrent neural networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 14, 2020</td>
<td></td>
<td>Recitation 2</td>
<td>Neural Networks Review</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Tuesday, February 18, 2020</td>
<td></td>
<td></td>
<td>(Holiday - President’s Day)</td>
</tr>
<tr>
<td></td>
<td>Thursday, February 20, 2020</td>
<td></td>
<td></td>
<td>Lecture 5</td>
<td>ML model interpretation I (SIS) (Brandon Carter Guest Lecture)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 21, 2020</td>
<td></td>
<td>Recitation 3</td>
<td>Interpreting ML models</td>
</tr>
<tr>
<td>PS2: TF Binding, ChIP, Motifs (out: Fri 2/21, Due: Fri 3/13)</td>
<td>4</td>
<td>Tuesday, February 25, 2020</td>
<td>Module 2: Chromatin structure / Model selection and uncertainty</td>
<td>Lecture 6</td>
<td>Chromatin accessibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, February 27, 2020</td>
<td></td>
<td>Lecture 7</td>
<td>Protein-DNA interactions and ChIP seq motif discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, February 28, 2020</td>
<td></td>
<td>Lecture 8</td>
<td>Model uncertainty and experiment design</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Tuesday, March 3, 2020</td>
<td></td>
<td>Lecture 9</td>
<td>Generative models (gradients, VAEs, GANs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 5, 2020</td>
<td></td>
<td>Recitation 4</td>
<td>Model uncertainty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 6, 2020</td>
<td></td>
<td>Lecture 10</td>
<td>Chromatin interactions and 3D genome organization</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Tuesday, March 10, 2020</td>
<td></td>
<td>Lecture 11</td>
<td>Dimensionality reduction (PCA, t-SNE, autoencoders)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 12, 2020</td>
<td></td>
<td>Recitation 5</td>
<td>Regulatory element models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 13, 2020</td>
<td></td>
<td>Lecture 12</td>
<td>The expressed genome and RNA splicing (RNA-seq)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Tuesday, March 17, 2020</td>
<td></td>
<td>Lecture 13</td>
<td>Quiz 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, March 19, 2020</td>
<td></td>
<td>Recitation 7</td>
<td>No recitation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, March 20, 2020</td>
<td></td>
<td></td>
<td>(Spring Vacation)</td>
</tr>
<tr>
<td>PS3: scRNA-seq tSNE analysis (out: Thu 3/12, due Fri 4/3)</td>
<td>8</td>
<td>Tuesday, March 31, 2020</td>
<td>Module 3: Expressed Genome / Dimensionality reduction</td>
<td>Lecture 14</td>
<td>scRNA-seq and cell labeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 2, 2020</td>
<td></td>
<td>Lecture 15</td>
<td>Manifolds, manifold mapping, word2vec</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 3, 2020</td>
<td></td>
<td>Lecture 8</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Tuesday, April 7, 2020</td>
<td></td>
<td>Recitation 8</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 9, 2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 10, 2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Tuesday, April 14, 2020</td>
<td>Module 4: Human Genetics - Genotype -&gt; Phenotype</td>
<td>Lecture 16</td>
<td>Deep learning in Disease Studies and Human Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 16, 2020</td>
<td></td>
<td>Lecture 17</td>
<td>eQTL prediction and variant prioritization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 17, 2020</td>
<td></td>
<td>Recitation 9</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Tuesday, April 21, 2020</td>
<td></td>
<td>Lecture 18</td>
<td>STARR-seq and GWAS studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 23, 2020</td>
<td></td>
<td>Lecture 19</td>
<td>High-throughput experimentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, April 24, 2020</td>
<td></td>
<td>Recitation 10</td>
<td>Protein structure prediction</td>
</tr>
<tr>
<td>No other psets</td>
<td>12</td>
<td>Tuesday, April 28, 2020</td>
<td>Module 5: Therapeutics and Diagnostics</td>
<td>Lecture 20</td>
<td>Therapeutic design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, April 30, 2020</td>
<td></td>
<td>Lecture 21</td>
<td>Imaging and genotype to phenotype (Guest: Adrian Dalca)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, May 5, 2020</td>
<td></td>
<td>Recitation 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Tuesday, May 5, 2020</td>
<td></td>
<td>Lecture 22</td>
<td>Quiz 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, May 7, 2020</td>
<td></td>
<td>Lecture 23</td>
<td>How to write, how to present</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friday, May 12, 2020</td>
<td></td>
<td>Recitation 12</td>
<td>Project work</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Tuesday, May 12, 2020</td>
<td>Projects</td>
<td>Lecture 24</td>
<td>Project Presentations I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, May 14, 2020</td>
<td></td>
<td>Lecture 25</td>
<td>Project Presentations II</td>
</tr>
</tbody>
</table>
Your programming environment

Problem 2

In this problem, we wish to use CNN to learn the motif of CTCF from sequences with similar di-nucleotide frequency. The positive samples are 101bp sequences centered at CTCF ChIP-seq peaks from GM12878 cell line. The negative sequences are generated by permuting the nucleotides in the positive sequences while keeping the di-nucleotide frequency.

We will provide functions for loading data, training and testing. You will:

- implement a CNN model with given specifications
- specify the initialization of parameters in the model
- train the model and evaluate on the test set

All the places where you need to fill in begins with "TODO" and ends with "END OF YOUR CODE".

```python
In [1]:
import tensorflow as tf, sys, numpy as np, h5py
from os.path import join, dirname, basename, exists, realpath
from os import makedirs
from tensorflow.examples.tutorials.mnist import input_data
from sklearn.metrics import roc_auc_score

In [2]:
data_folder = '../data/motif_disc'
batch_size = 128
valid_size = 2000
epochs = 20
best_model_file = join('../output', basename(data_folder), 'best_model.ckpt')
if not exists(dirname(best_model_file)):
    makedirs(dirname(best_model_file))

In [3]:
# Function to load the data embedded in the previous problem and their labels
def load_data(mdir):
    train = h5py.File(join(mdir, 'train.h5'), 'r')
```
Your computing resource
Cloud Platform Education Grants

Use credits provided to you via the Google Cloud Platform Education Grants program to access Google Cloud Platform. Get what you need to build and run your apps, websites and services.

Thank you for your interest in Google Cloud Platform Education Grants. Please fill out the form below to receive a coupon code for credit to use on Google Cloud Platform.

First Name

Last Name

School Email

@mit.edu

If you do not see your domain listed, please contact your course instructor: gifford@mit.edu

By clicking "Submit" below, you agree that we may share the following information with your educational institution and course instructor (gifford@mit.edu): (1) personal information that you provide to us on this form and (2) information regarding your use of the coupon and Google Cloud Platform products.

Submit
What is Machine Learning?

[Shalev-Shwartz and Ben-David, 2014]:
   “Learning is the process of converting experience into expertise or knowledge.”
What is Machine Learning?

[Shalev-Shwartz and Ben-David, 2014]:

“Learning is the process of converting experience into expertise or knowledge.”

[Mohri et al., 2012]:

“Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions.”
What is Machine Learning?

[Shalev-Shwartz and Ben-David, 2014]:
“Learning is the process of converting experience into expertise or knowledge.”

[Mohri et al., 2012]:
“Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions.”

[Murphy, 2012]:
“The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.”
What is Machine Learning?

[Shalev-Shwartz and Ben-David, 2014]:
“Learning is the process of converting experience into expertise or knowledge.”

[Mohri et al., 2012]:
“Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions.”

[Murphy, 2012]:
“The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.”

[Hastie et al., 2001]:
“[...] state the learning task as follows: given the value of an input vector $\mathbf{x}$, make a good prediction of the output $\mathbf{y}$, denoted by $\hat{\mathbf{y}}$.”
What is Machine Learning?

A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.

[Mitchell, 1997]
What is Machine Learning?

A computer program is said to learn from **experience E**

with respect to some **class of tasks T**

and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E.

[Mitchell, 1997]

Problem Set 1

- experience E: training set of images of handwritten digits with labels (training set)
- task T: classifying handwritten digits within new images (test set)
- performance measure P: percent of test set digits correctly classified in new images (test set)
What is Machine Learning?

A computer program is said to learn from
experience $E$

with respect to some
class of tasks $T$

and
performance measure $P$,

if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.

[Mitchell, 1997]
What is Machine Learning?

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

[Mitchell, 1997]

**Problem Set 1**

- experience E: training set of images of handwritten digits with labels (training set)
- task T: classifying handwritten digits within new images (test set)
- performance measure P: percent of test set digits correctly classified in new images (test set)
Notation

\( a, b, c_i \)  scalar (slanted, lower-case)

\( \mathbf{a}, \mathbf{b}, \mathbf{c} \)  vector (bold, slanted, lower-case)

\( \mathbf{A}, \mathbf{B}, \mathbf{C} \)  matrix (bold, slanted, upper-case)

\( \mathcal{A}, \mathcal{B}, \mathcal{C} \)  tensor (bold, upright, upper-case)

\( \mathcal{A}, \mathcal{B}, \mathcal{C} \)  set (calligraphic, slanted, upper-case)
Notation

$a, b, c_i$ scalar (slanted, lower-case)

$a, b, c$ vector (bold, slanted, lower-case)

$A, B, C$ matrix (bold, slanted, upper-case)

$A, B, C$ tensor (bold, upright, upper-case)

$\mathcal{A}, \mathcal{B}, \mathcal{C}$ set (calligraphic, slanted, upper-case)

$\mathcal{X}$ input space or feature space

$\mathbf{X}, \mathbf{X}$ dataset example matrix or tensor

$x^{(i)}$ $i$th example of dataset, one row of $\mathbf{X}$

$x_j^{(i)}, x_j$ feature $j$ of example $x^{(i)}$

$\mathcal{Y}$ label space

$y^{(i)}$ label of example $i$

$\hat{y}^{(i)}$ predicted label of example $i$
Terminology

Input $X \in \mathcal{X}$:
- **features** (in machine learning)
- predictors (in statistics)
- independent variables (in statistics)
- regressors (in regression models)
- input variables
- covariates

Output $y \in \mathcal{Y}$:
- **labels** (in machine learning)
- responses (in statistics)
- dependent variables (in statistics)
- regressand (in regression models)
- target variables

Training set $S_{\text{training}} = \{(X^{(i)}, y^{(i)})\}_{i=1}^{N} \in \{\mathcal{X}, \mathcal{Y}\}^N$, where $N$ is number of training examples

An example is a collection of features (and an associated label)

Training: use $S_{\text{training}}$ to learn functional relationship $f : \mathcal{X} \to \mathcal{Y}$
Terminology

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]
\[ f(x; \theta) = \hat{y} \]

\[ \theta : \]
- **weights and biases** (intercepts)
- coefficients \( \beta \)
- parameters

\[ f : \]
- model
- hypothesis \( h \)
- classifier
- predictor
- discriminative models: \( P(Y|X) \)
- generative models: \( P(X, Y) \)

Problem Set 1

\[ x \in [0, 1]^{784} \]
\[ \hat{y} \in [0, 1]^{10} \]
\[ W \in \mathbb{R}^{784 \times 10} \]
\[ b \in \mathbb{R}^{10} \]
\[ f(x; W, b) = \phi_{\text{softmax}}(W^T x + b) \]
Problem Set 1

input space:
\[ \mathcal{X} = \{0, 1, \ldots, 255\}^{28 \times 28} \]

after rescaling:
\[ \mathcal{X}' = [0, 1]^{28 \times 28} \]

after flattening:
\[ \mathcal{X}'' = [0, 1]^{784} \]
Problem Set 1

input space:
\[ \mathcal{X} = \{0, 1, \ldots, 255\}^{28 \times 28} \]

after rescaling:
\[ \mathcal{X}' = [0, 1]^{28 \times 28} \]

after flattening:
\[ \mathcal{X}'' = [0, 1]^{784} \]

integer-encoded label space:
\[ \mathcal{Y}_i = \{0, 1, \ldots, 9\} \]

one-hot-encoded label space:
\[ \mathcal{Y}_h = [0, 1]^{10} \]
Types of Machine Learning

**Classification**

$\mathcal{Y} \neq \emptyset$

- $\mathcal{Y} = \mathbb{R}$: regression
- $\mathcal{Y} = \mathbb{R}^K, K > 1$: multivariate regression
- $\mathcal{Y} = \{0, 1\}$: binary classification
- $\mathcal{Y} = \{1, \ldots, K\}$: multi-class classification (integer encoding)
- $\mathcal{Y} = \{0, 1\}^K, K > 1$: multi-label classification

$\mathcal{Y} = \emptyset$: unsupervised learning

---

**Regression**

$\mathcal{Y} \neq \emptyset$

- $\mathcal{Y} = \mathbb{R}$: regression

---

**Unsupervised learning**

$\mathcal{Y} = \emptyset$: unsupervised learning
Types of Machine Learning

Problem Set 1

- task: every $\mathbf{X}$ has an associated $\mathbf{y} \implies$ supervised learning
- subtask: $\mathcal{Y} = \{0, \ldots, 9\} \implies$ multi-class classification
- method: we use softmax regression (also known as multinomial logistic regression) as multi-class classification method
Objective functions

An **objective function** $J(\Theta)$ is the function that you optimize when training machine learning models. It is usually in the form of (but not limited to) one or combinations of the following:

**Loss / cost / error function $L(\hat{y}, y)$:**

- **Classification**
  - 0-1 loss
  - cross-entropy loss
  - hinge loss

- **Regression**
  - mean squared error (MSE, $L_2$ norm)
  - mean absolute error (MAE, $L_1$ norm)
  - Huber loss (hybrid between $L_1$ and $L_2$ norm)

- **Probabilistic inference**
  - Kullback-Leibler divergence (KL divergence)

**Likelihood function / posterior:**

- negative log-likelihood (NLL) in maximum likelihood estimation (MLE)
- posterior in maximum a posteriori estimation (MAP)

**Regularizers and constraints**

- $L_1$ regularization $||\Theta||_1 = \lambda \sum_{i=1}^{N} |\theta_i|$
- $L_2$ regularization $||\Theta||_2^2 = \lambda \sum_{i=1}^{N} \theta_i^2$
- max-norm $||\Theta||_2^2 \leq c$
Loss functions for classification

0-1 loss:

\[
\mathcal{L}_{0-1}(\hat{y}, y) = \sum_{i=1}^{N} \mathbb{1}(\lfloor \hat{y}^{(i)} \rfloor \neq y^{(i)}) = \sum_{i=1}^{N} \begin{cases} 
1, & \text{for } \hat{y}^{(i)} \neq y^{(i)} \\
0, & \text{for } \hat{y}^{(i)} = y^{(i)}
\end{cases}
\]

where \([x]\) is the function that rounds \(x\) to the nearest integer.
Loss functions for classification

0-1 loss:

\[ \mathcal{L}_{0-1}(\hat{y}, y) = \sum_{i=1}^{N} \mathbb{I}(\lceil \hat{y}^{(i)} \rceil \neq y^{(i)}) = \sum_{i=1}^{N} \begin{cases} 1, & \text{for } \hat{y}^{(i)} \neq y^{(i)} \\ 0, & \text{for } \hat{y}^{(i)} = y^{(i)} \end{cases} \]

where \([x]\) is the function that rounds \(x\) to the nearest integer.

Binary cross-entropy loss (for binary classification):

\[ \mathcal{L}_{\text{BCE}} = \text{NLL (Negative Log Likelihood)} \]
Likelihood is defined using the Bernoulli distribution

\[ p(\hat{y}^{(i)}, y^{(i)}) = (\hat{y}^{(i)})^{y^{(i)}}(1 - \hat{y}^{(i)})^{1 - y^{(i)}} \]
Loss functions for classification

0-1 loss:

\[ L_{0-1}(\hat{y}, y) = \sum_{i=1}^{N} \mathbb{I}([\hat{y}^{(i)}] \neq y^{(i)}) = \sum_{i=1}^{N} \begin{cases} 1, & \text{for } \hat{y}^{(i)} \neq y^{(i)} \\ 0, & \text{for } \hat{y}^{(i)} = y^{(i)} \end{cases} \]

where \([x]\) is the function that rounds \(x\) to the nearest integer.

Binary cross-entropy loss (for binary classification):

\[ L_{\text{BCE}} = \text{NLL (Negative Log Likelihood)} \]

Likelihood is defined using the Bernoulli distribution

\[ p(\hat{y}^{(i)}, y^{(i)}) = (\hat{y}^{(i)})y^{(i)}(1 - \hat{y}^{(i)})(1 - y^{(i)}) \]

\[ L_{\text{BCE}}(\hat{y}, y) = \sum_{i=1}^{N} -y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \]

\[ = \sum_{i=1}^{N} \begin{cases} - \log(\hat{y}^{(i)}), & \text{for } y^{(i)} = 1 \\ - \log(1 - \hat{y}^{(i)}), & \text{for } y^{(i)} = 0 \end{cases} \]
Loss functions for classification

**Binary cross-entropy loss** (for binary classification):

\[
\mathcal{L}_{\text{BCE}}(\hat{y}, y) = \sum_{i=1}^{N} -y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})
\]

\[
= \sum_{i=1}^{N} \begin{cases} 
- \log(\hat{y}^{(i)}), & \text{for } y^{(i)} = 1 \\
- \log(1 - \hat{y}^{(i)}), & \text{for } y^{(i)} = 0
\end{cases}
\]

<table>
<thead>
<tr>
<th>$y$</th>
<th>$\hat{y}$</th>
<th>$[\hat{y}]$</th>
<th>$\mathcal{L}_{0-1}(\hat{y}, y)$</th>
<th>$\mathcal{L}_{\text{BCE}}(\hat{y}, y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 0, 0]</td>
<td>[0.9, 0.2, 0.4]</td>
<td>[1, 0, 0]</td>
<td>0</td>
<td>0.84</td>
</tr>
<tr>
<td>[1, 1, 0]</td>
<td>[0.6, 0.4, 0.1]</td>
<td>[1, 0, 0]</td>
<td>1</td>
<td>1.53</td>
</tr>
<tr>
<td>[1, 0, 1]</td>
<td>[0.1, 0.7, 0.3]</td>
<td>[0, 1, 0]</td>
<td>3</td>
<td>4.71</td>
</tr>
</tbody>
</table>
Problem Set 1

**Categorical cross-entropy loss** (for multi-class classification with $K$ classes):

$$
\mathcal{L}_{\text{CCE}}(\hat{y}, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_j^{(i)} \log(\hat{y}_j^{(i)}),
$$

where

$$
\hat{y}_j^{(i)} = \frac{\exp(z_j^{(i)})}{\sum_{k=1}^{K} \exp(z_k^{(i)})}
$$

if softmax is used

note: $y_j^{(i)} = 1$ only if $x^{(i)}$ belongs to class $j$ and otherwise $y_j^{(i)} = 0$

Probabilistic interpretation:

$\mathcal{L}_{\text{CCE}} = \text{NLL}$, if likelihood is defined using the categorical distribution
Loss functions for regression

Mean squared error:

$$\mathcal{L}_{\text{MSE}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$$

Probabilistic interpretation:

$\mathcal{L}_{\text{MSE}} = \text{NLL}$, under the assumption that the noise is normally distributed, with constant mean and variance
Loss functions for regression

Mean squared error:

\[
\mathcal{L}_{\text{MSE}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2
\]

Probabilistic interpretation:
\(\mathcal{L}_{\text{MSE}} = \text{NLL}\), under the assumption that the noise is normally distributed, with constant mean and variance

Mean absolute error:

\[
\mathcal{L}_{\text{MAE}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} |y^{(i)} - \hat{y}^{(i)}|
\]
Loss functions for regression

Mean squared error:

\[ \mathcal{L}_{\text{MSE}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2 \]

Probabilistic interpretation:
\( \mathcal{L}_{\text{MSE}} = \text{NLL} \), under the assumption that the noise is normally distributed, with constant mean and variance

Mean absolute error:

\[ \mathcal{L}_{\text{MAE}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} |y^{(i)} - \hat{y}^{(i)}| \]

\[
\begin{array}{cccc}
 y & \hat{y} & \mathcal{L}_{\text{MSE}}(\hat{y}, y) & \mathcal{L}_{\text{MAE}}(\hat{y}, y) \\
[3.2, 1.2, 0.3] & [3.1, 1.3, 0.4] & 0.01 & 0.1 \\
[2.1, 0.1, -5.1] & [2.0, -0.1, 1.2] & 13.25 & 2.2 \\
[-0.1, 3.1, 0.5] & [0.1, 3.3, -0.5] & 0.36 & 0.47 \\
\end{array}
\]
Empirical risk minimization

Expected risk (loss) associated with hypothesis $h(x)$:

$$R_{\text{exp}}(h) = \mathbb{E}(L(h(x), y)) = \int_{\mathcal{X} \times \mathcal{Y}} L(h(x), y)p(x, y)dx dy$$

Minimize $R_{\text{exp}}(h)$ to find optimal hypothesis $h$:

$$h = \arg\min_{h \in \mathcal{F}} R_{\text{exp}}(h)$$

Problem:
- distribution $p(x, y)$ unknown
- $\mathcal{F}$ is too large (set of all functions from $\mathcal{X}$ to $\mathcal{Y}$)
Empirical risk minimization

**Empirical risk** associated with hypothesis $h(x)$:

$$R_{\text{emp}}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(h(x^{(i)}), y^{(i)})$$

Minimize $R_{\text{emp}}(h)$ to find $\hat{h}$:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} R_{\text{emp}}(h)$$

In practice:

- instead of $p(x, y)$, we use training set $\mathcal{S}_{\text{training}}$
- instead of $\mathcal{F}$, we use $\mathcal{H} \subset \mathcal{F}$, e.g., all polynomials of degree 5
Optimizing objective function

Gradient descent

- initialize model parameters $\theta_0, \theta_1, \ldots, \theta_m$
- repeat until converge, for all $\theta_i$

$$\theta_i^t \leftarrow \theta_i^{t-1} - \lambda \frac{\partial}{\partial \theta_i^{t-1}} J(\Theta),$$

where the objective function $J(\Theta)$ is evaluated over all training data $\{(X^{(i)}, y^{(i)})\}_{i=1}^N$.

Problem Set 1

**Stochastic Gradient Descent (SGD):** in each step, randomly sample a mini-batch from the training data and update the parameters using gradients calculated from the mini-batch only.
Training, validation, test sets

**Training set** ($S_{\text{training}}$):
- set of examples used for learning
- usually 60 - 80 % of the data

**Validation set** ($S_{\text{validation}}$):
- set of examples used to tune the model hyperparameters
- usually 10 - 20 % of the data

**Test set** ($S_{\text{test}}$):
- set of examples used only to assess the performance of fully-trained model
- after assessing test set performance, model must not be tuned further
- usually 10 - 30 % of the data
## Confusion matrix and derived metrics

<table>
<thead>
<tr>
<th>Predicted condition</th>
<th>True condition</th>
<th>Accuracy</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total population</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Condition positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Condition negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted condition</td>
<td>True positive</td>
<td>Recall, Sensitivity</td>
<td></td>
</tr>
<tr>
<td>predicted</td>
<td>Power</td>
<td>Recall, Sensitivity = [ \frac{\sum \text{True positive}}{\sum \text{Condition positive}} ]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>False positive</td>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type I error</td>
<td>Specificity = [ \frac{\sum \text{True negative}}{\sum \text{Condition negative}} ]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>False negative</td>
<td>F_1 score</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type II error</td>
<td>F_1 score = [ 1 \cdot \frac{1}{\left( \frac{1}{\text{Recall}} + \frac{1}{\text{Precision}} \right)} ]</td>
<td></td>
</tr>
</tbody>
</table>

### Problem Set 1

**Accuracy**: proportion of true predictions - \[ \frac{(TP + TN)}{(TP + FP + TN + FN)} \]
Receiver Operating Characteristic (ROC) Performance

Area Under the ROC Curve (AuROC)

AuROC is a common metric for comparing classification methods

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Problematic when we have an unbalanced dataset (example more positives than negatives)
Precision Recall Curve (PRC) Performance

Area Under the PRC (AuPRC)

Precision = PPV = TP / (TP + FP) = 1 - FDR
Recall = TPR = TP / (TP + FN)

Useful when datasets are unbalanced
ROC and PRC curves are complementary

Recall

\[ FPR = \frac{FP}{FP + TN} \]
\[ \text{Precision} = PPV = \frac{TP}{TP + FP} = 1 - FDR \]
\[ \text{Recall} = TPR = \frac{TP}{TP + FN} \]
Regression Metric 1 - Pearson Correlation

Pearson correlation coefficient is $r$. $r^2$ is the fraction of linearly explained variance

$$r = \frac{(x-\bar{x}) \cdot (y-\bar{y})}{||x|| \cdot ||y||}$$
Regression Metric 2 - Spearman Rank Correlation

Pearson correlation of observation ranks
For ties assign fractional ranks by average rank in ascending order
Correlation significance tests

\[ t \text{ is distributed as Student's } t\text{-distribution with } n - 2 \text{ degrees of freedom under the null hypothesis} \]

\[ n \text{ is the number of observations} \]

\[ t = r \sqrt{\frac{n-2}{1-r^2}} \]

Alternatively we can permute values to observe the empirical distribution of null correlations.
Two sided tests are used when we are testing for a difference without regard to direction.

Two sided tests allocate half the area to each direction.
Thus they are more strict if you only wish to test in one direction.
Classifier significance test

Binomial test for probability that null model would produce observed results

- \( n \) is the number of observations in test set
- \( k \) is the number classified correctly test set
- \( p \) is the probability classifier will make correct choice at random

Probability that exactly \( k \) observations are classified correctly by null:

\[
Pr(x = k) = \binom{n}{k} p^k (1 - p)^{n - k}
\]
Classifier significance test

Binomial test for probability that null model would produce observed results

- \( n \) is the number of observations in test set
- \( k \) is the number classified correctly test set
- \( p \) is the probability classifier will make correct choice at random

Probability that exactly \( k \) observations are classified correctly by null:

\[
Pr(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}
\]

Test that \( k \) or greater would have been classified correctly by null:

\[
p = \sum_{i=k}^{n} Pr(x = i)
\]

This can be approximated by a Chi-squared test.
Multiple hypothesis correction is important

If we ask $m$ questions we need to adjust our probability that the null is likely

$$p_{\text{single}} \quad \text{Probability one test occurred by chance}$$
Multiple hypothesis correction is important

If we ask $m$ questions we need to adjust our probability that the null is likely

$p_{\text{single}}$ Probability one test occurred by chance

$p_{\text{corrected}} \leq m \times p_{\text{single}}$ from Boole’s inequality results in the Bonferonni correction
Multiple hypothesis correction is important

If we ask $m$ questions we need to adjust our probability that the null is likely

- $p_{\text{single}}$: Probability one test occurred by chance
- $p_{\text{corrected}} \leq m \times p_{\text{single}}$: from Boole’s inequality results in the Bonferonni correction
- $p_{\text{single}} \leq \frac{p_{\text{corrected}}}{m}$: Filter for significant events
### Multiple hypothesis correction is important

If we ask $m$ questions we need to adjust our probability that the null is likely

<table>
<thead>
<tr>
<th>$p_{\text{single}}$</th>
<th>Probability one test occurred by chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{corrected}} \leq m \cdot p_{\text{single}}$</td>
<td>from Boole’s inequality results in the Bonferonni correction</td>
</tr>
<tr>
<td>$p_{\text{single}} \leq \frac{p_{\text{corrected}}}{m}$</td>
<td>Filter for significant events</td>
</tr>
</tbody>
</table>

#### Benjamini-Hochberg uses a desired false discover rate to provide a relaxed bound

- $\alpha$ is our desired false discovery rate (FDR)
- $m$ is the number of tests $H_1 \ldots H_m$
- $P_1 \ldots P_m$ are their p-values in ascending order
  - Find the largest $k$ such that $P_k \leq \frac{k}{m} \alpha$
  - Reject the null hypothesis for all $H_i$ for $i = 1, \ldots, k$

Single test p-values are 0.003, 0.006, 0.020, 0.045, 0.600
Multiple hypothesis correction is important

If we ask $m$ questions we need to adjust our probability that the null is likely

\[ p_{single} \] Probability one test occurred by chance

\[ p_{corrected} \leq m \times p_{single} \] from Boole’s inequality results in the Bonferonni correction

\[ p_{single} \leq \frac{p_{corrected}}{m} \] Filter for significant events

Benjamini-Hochberg uses a desired false discover rate to provide a relaxed bound

\[ \alpha \] is our desired false discovery rate (FDR)

\[ m \] is the number of tests $H_1 \ldots H_m$

\[ P_1 \ldots P_m \] are their p-values in ascending order

- Find the largest $k$ such that $P_k \leq \frac{k}{m} \alpha$
- Reject the null hypothesis for all $H_i$ for $i = 1, \ldots, k$

Which transcription factors $TF_1 \ldots TF_5$ bind with a corrected significance of .05?

Single test p-values are 0.003, 0.006, 0.020, 0.045, 0.600
Correlation is not causation
The Datasaurus Dozen - J. Matejka, G. Fitzmaurice

X Mean: 54.26
Y Mean: 47.83
X SD: 16.76
Y SD: 26.93
Corr.: -0.06
Quo vadis, 6.874?

- neural networks (NNs)
  - convolutional neural networks (CNNs)
  - recurrent neural networks (RNNs)
  - residual neural networks
  - (variational) autoencoders (VAEs)
  - generative adversarial networks (GANs)
- regularization
  - $L_1$ regularization
  - $L_2$ regularization
  - dropout
  - early stopping
- model selection
  - cross-validation (CV)
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
- model interpretation methods
  - sufficient input subsets (SIS)
  - saliency maps
- model uncertainty
  - identifying out of distribution inputs
  - ensembles and calibrated uncertainty
- dimensionality reduction methods
  - principal component analysis (PCA)
  - t-SNE
  - autoencoders
  - non-negative matrix factorization (NMF)
- hyperparameter optimization and AutoML
*The Elements of Statistical Learning.*
Springer, New York, NY, USA.

*Machine Learning.*

*Foundations of Machine Learning.*
MIT Press, Cambridge, MA, USA.

Murphy, K. P. (2012).
*Machine learning : a probabilistic perspective.*
MIT Press, Cambridge, MA, USA.

Shalev-Shwartz, S. and Ben-David, S. (2014).
*Understanding Machine Learning: From Theory to Algorithms.*
Cambridge University Press, New York, NY, USA.