Recall: Binary Classification

- Logistic regression: \( P(y = 1|x) = \frac{\exp \left( \sum_{i=1}^{n} w_i x_i \right)}{1 + \exp \left( \sum_{i=1}^{n} w_i x_i \right)} \)

- Decision rule: \( P(y = 1|x) \geq 0.5 \Leftrightarrow w^\top x \geq 0 \)

- Gradient (unregularized): \( x(y - P(y = 1|x)) \)

- SVM: quadratic program to minimize weight vector norm \( w \) with slack

- Decision rule: \( w^\top x \geq 0 \)

- (Sub)gradient (unregularized): 0 if correct with margin of 1, else \( x(2y - 1) \)

Loss Functions

- Hinge (SVM)
- 0-1 (ideal)
- Perceptron
- Logistic
Sentiment Analysis

- Classify sentence as positive or negative sentiment
  - Positive: this movie was great! would watch again
  - Negative: the movie was gross and overwrought, but I liked it
  - this movie was not really very enjoyable

- Bag-of-words doesn’t seem sufficient (discourse structure, negation)
- There are some ways around this: extract bigram feature for “not X” for all X following the not

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan (2002)
### Text Classification

- **A Cancer Conundrum: Too Many Drug Trials, Too Few Patients**
  
  Breakthroughs in immunotherapy and a rush to develop profitable new treatments have brought a crush of clinical trials scrambling for patients.
  
  By DINA KELATA

- **Yankees and Mets Are on Opposite Tracks This Subway Series**
  
  As they meet for a four-game series, the Yankees are playing for a postseason spot, and the most the Mets can hope for is to play spoiler.
  
  By FILIP BENDY

- ~20 classes

### Image Classification

- **Health**
  - Dog

- **Sports**
  - Car

- Thousands of classes (ImageNet)

### Entity Linking

- **Although he originally won the event, the United States Anti-Doping Agency announced in August 2012 that they had disqualified Armstrong from his seven consecutive Tour de France wins from 1999–2005.**

  - Lance Edward Armstrong is an American former professional road cyclist
  - Armstrong County is a county in Pennsylvania...

- 4,500,000 classes (all articles in Wikipedia)

### Reading Comprehension

- One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn’t pay, and instead headed home.

  1) Where did James go after he went to the grocery store?
  A) his deck
  B) his freezer
  C) a fast food restaurant
  D) his room

  After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

- Multiple choice questions, 4 classes (but classes change per example)

Richardson (2013)
Binary Classification

- Binary classification: one weight vector defines positive and negative classes

Multiclass Classification

- Can we just use binary classifiers here?

Multiclass Classification

- One-vs-all: train $k$ classifiers, one to distinguish each class from all the rest
- How do we reconcile multiple positive predictions? Highest score?

Multiclass Classification

- Not all classes may even be separable using this approach

Slide credit: Vivek Srikumar
Multiclass Classification

- All-vs-all: train \( \frac{n(n-1)}{2} \) classifiers to differentiate each pair of classes
- Binary classification: one weight vector defines both classes
- Multiclass classification: one weight vector per class, decision is \( \text{argmax} \)

Again, how to reconcile?

Decision rule:

- Can also have one weight vector per class: \( \text{argmax}_{y \in \mathcal{Y}} w_y^T f(x) \)
- Why do we do with separate feature vectors? Let’s see!

Formally: instead of two labels, we have an output space \( \mathcal{Y} \) containing a number of possible classes
- Same machinery that we’ll use later for exponentially large output spaces, including sequences and trees
- Decision rule: \( \text{argmax}_{y \in \mathcal{Y}} w^T f(x, y) \)
- Multiple feature vectors, one weight vector

Block Feature Vectors

- Equivalent to having three weight vectors, but this formulation is more general if the features depend on \( y \)
- Base feature function:
  \[
  f(x, y) = \begin{cases} \text{[contains drug]}, \text{[contains patients]}, \text{[contains baseball]} = [1, 1, 0] & \text{if label} = \text{Health} \\ [0, 0, 0, 1, 0, 0, 0] & \text{if label} = \text{Sports} \end{cases}
  \]
- Too many drug trials, too few patients

Decision rule: \( \text{argmax}_{y \in \mathcal{Y}} w^T f(x, y) \)
### Making Decisions

*too many drug trials, too few patients*

- Health
- Sports
- Science

\[
f(x) = \begin{cases} 1, & \text{contains drug} \\ 0, & \text{contains patients} \\ 0, & \text{contains baseball} \end{cases}
\]

\[
f(x, y = \text{Health}) = [1, 0, 0; 0, 0, 0]
\]

\[
f(x, y = \text{Sports}) = [0, 0, 1; 0, 0, 0]
\]

\[
w = [+2.1, +2.3, -5, -2.1, -3.8, +5.2, +1.1, -1.7, -1.3]
\]

\[
w^T f(x, y) = \begin{cases} \text{Health: +4.4} \\ \text{Sports: -5.9} \\ \text{Science: -1.9} \end{cases}
\]

### Multiclass Logistic Regression

\[
P(y|x) = \frac{\exp \left( w^T f(x, y) \right)}{\sum_{y' \in Y} \exp \left( w^T f(x, y') \right)}
\]

- Compare to binary:
  \[
P(y = 1|x) = \frac{\exp(w^T f(x))}{1 + \exp(w^T f(x))}
\]

  sum over output space to normalize

- negative class implicitly had \( f(x, y = 0) = \) the zero vector

- Training: maximize \( \mathcal{L}(x, y) = \sum_{j=1}^{n} \log P(y_j| x_j) \)
  \[
  = \sum_{j=1}^{n} \left( w^T f(x_j, y_j^*) - \log \sum_y \exp(w^T f(x_j, y)) \right)
  \]

### Training

- Multiclass logistic regression \( P(y|x) = \frac{\exp \left( w^T f(x, y) \right)}{\sum_{y' \in Y} \exp \left( w^T f(x, y') \right)} \)

- Likelihood \( \mathcal{L}(x, y_j^*) = w^T f(x_j, y_j^*) - \log \sum_y \exp(w^T f(x_j, y)) \)

\[
\frac{\partial}{\partial w_i} \mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \sum_y f_i(x_j, y) \exp(w^T f(x_j, y))
\]

\[
\frac{\partial}{\partial w_i} \mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \sum_y f_i(x_j, y) P(y|x_j)
\]

\[
\frac{\partial}{\partial w_i} \mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \mathbb{E}_y[f_i(x_j, y)]
\]

"towards gold feature value, away from expectation of feature value"

### Logistic Regression: Summary

- Model: \( P(y|x) = \frac{\exp \left( w^T f(x, y) \right)}{\sum_{y' \in Y} \exp \left( w^T f(x, y') \right)} \)

- Inference: \( \text{argmax}_y P(y|x) \)

- Learning: gradient ascent on the discriminative log-likelihood

\[
\frac{\partial}{\partial w_i} \mathcal{L}(x_j, y_j^*) = f_i(x_j, y_j^*) - \mathbb{E}_y[f_i(x_j, y)]
\]

"towards gold feature value, away from expectation of feature value"
Training

- Are all decisions equally costly?
  
  - too many drug trials, too few patients

  - Predicted Sports: bad error
  - Predicted Science: not so bad

- We can define a loss function $\ell(y, y^*)$
  
  $\ell(\text{Sports}, \text{Health}) = 3$
  $\ell(\text{Science}, \text{Health}) = 1$

Multiclass SVM

Minimize $\lambda\|w\|^2 + \sum_{j=1}^m \xi_j$

s.t. $\forall j \xi_j \geq 0$

$\forall j \forall y \in \mathcal{Y} \ w^T f(x_j, y_j^*) \geq w^T f(x_j, y) + \ell(y, y_j^*) - \xi_j$

Correct prediction now has to beat every other class

Score comparison is more explicit now

The 1 that was here is replaced by a loss function

Multiclass SVM

Minimize $\lambda\|w\|^2 + \sum_{j=1}^m \xi_j$

s.t. $\forall j \xi_j \geq 0$

$\forall j (2y_j - 1)(w^T x_j) \geq 1 - \xi_j$

$\forall j \forall y \in \mathcal{Y} \ w^T f(x_j, y_j^*) \geq w^T f(x_j, y) + \ell(y, y_j^*) - \xi_j$

How does this quantification come into play?

- One slack variable per example, so it’s set to be whatever the most violated constraint is for that example

$\xi_j = \max_{y \in \mathcal{Y}} [w^T f(x_j, y) + \ell(y, y_j^*) - w^T f(x_j, y_j^*) - \xi_j$

Plug in the gold $y$ and you get 0, so slack is always nonnegative!

Loss-Augmented Decoding

$\xi_j = \max_{y \in \mathcal{Y}} [w^T f(x_j, y) + \ell(y, y_j^*) - w^T f(x_j, y_j^*)$

too many drug trials, too few patients

<table>
<thead>
<tr>
<th></th>
<th>w^T f(x, y)</th>
<th>Loss</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>+2.4</td>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td>Sports</td>
<td>+1.3</td>
<td>3</td>
<td>4.3</td>
</tr>
<tr>
<td>Science</td>
<td>+1.8</td>
<td>1</td>
<td>2.8</td>
</tr>
</tbody>
</table>

- Sports is most violated constraint, slack = 4.3 — 2.4 = 1.9

- Perceptron would make no update, regular SVM would pick Science
Computing the Subgradient

Minimize \( \lambda \|w\|^2 + \sum_{j=1}^{m} \xi_j \)

s.t. \( \forall j \xi_j \geq 0 \)

\( \forall j \forall y \in Y \ w^T f(x_j, y_j) \geq w^T f(x_j, y) + \ell(y, y_j^\star) - \xi_j \)

- If \( \xi_j = 0 \), the example is not a support vector, gradient is zero
- Otherwise, \( \xi_j = \max_{y \in Y} w^T f(x_j, y) + \ell(y, y_j^\star) - w^T f(x_j, y_j^\star) \)

\( \frac{\partial}{\partial w_i} \xi_j = f_i(x_j, y_{\max}) - f_i(x_j, y_j^\star) \) \( \implies \) (update looks backwards — we’re minimizing here!)

- Perceptron-like, but we update away from *loss-augmented* prediction

Softmax Margin

- Can we include a loss function in logistic regression?

\[
P(y|x) = \frac{\exp \left( w^T f(x, y) + \ell(y, y_j^\star) \right)}{\sum_{y'} \exp \left( w^T f(x, y') + \ell(y', y_j^\star) \right)}
\]

- Likelihood is artificially higher for things with high loss — training needs to work even harder to maximize the likelihood of the right thing!

- Biased estimator for original likelihood, but better loss

Entity Linking

Although he originally won the event, the United States Anti-Doping Agency announced in August 2012 that they had disqualified Armstrong from his seven consecutive Tour de France wins from 1999–2005.

- 4.5M classes, not enough data to learn features like “Tour de France <-> en/wiki/Lance_Armstrong”
- Instead, features \( f(x, y) \) look at the actual article associated with \( y \)

- Although he originally won the event, the United States Anti-Doping Agency announced in August 2012 that they had disqualified Armstrong from his seven consecutive Tour de France wins from 1999–2005.

  - \( \text{tf-idf}(\text{doc}, w) = \text{freq of } w \text{ in } \text{doc} * \log(4.5M/# \text{Wiki articles } w \text{ occurs in}) \)
  - \( \text{the}: \text{occurs in every article, } \text{tf-idf} = 0 \)
  - \( \text{cyclist}: \text{occurs in } 1\% \text{ of articles, } \text{tf-idf} = \# \text{ occurrences} * \log10(100) \)
  - \( \text{tf-idf}(\text{doc}) = \text{vector of } \text{tf-idf}(\text{doc}, w) \text{ for all words in vocabulary (50,000)} \)
  - \( f(x, y) = [\cos(\text{tf-idf}(x), \text{tf-idf}(y)), \ldots \text{ other features}] \)
Structured Prediction

- Four elements of a structured machine learning method:
  - Model: probabilistic, max-margin, deep neural network
  - Objective
  - Inference: just maxes so far, but will get harder
  - Training: gradient descent

Optimization

- Stochastic gradient ascent
  \[ w \leftarrow w + \alpha g, \quad g = \frac{\partial}{\partial w} L \]
  - Very simple to code up
  - "First-order" technique: only relies on having gradient
  - Difficult to tune step size
- Newton’s method
  - Second-order technique
  - Optimizes quadratic instantly
  \[ w \leftarrow w + \left( \frac{\partial^2}{\partial w^2} L \right)^{-1} g \]
  - Inverse Hessian: \( n \times n \) mat, expensive!
- Quasi-Newton methods: L-BFGS, etc.
  - Approximate inverse Hessian with gradients over time

AdaGrad

- Optimized for problems with sparse features
  - Per-parameter learning rate: smaller updates are made to parameters that get updated frequently
  \[ w_i \leftarrow w_i + \alpha \frac{1}{\sum_{t=1}^{T} g_{t,i}^2} g_{t,i} \]
  - accumulate sum of squared gradients from previous updates
- Generally much more robust, requires little tuning of learning rates
- Other techniques for optimizing deep models — more later!

Structured Prediction

- Design tradeoffs need to reflect interactions:
  - Model and objective are coupled: probabilistic model <-> maximize likelihood
  - ...but not always: a linear model or neural network can be trained to minimize any differentiable loss function
  - Inference governs what learning: need to be able to compute expectations to use logistic regression

Duchi et al. (2011)
Summary

- You’ve now seen everything you need to implement multi-class classification models

- Next time: HMMs (POS tagging)

- In 2 lectures: CRFs (NER)