Maximum Entropy Models/
Logistic Regression

CMSC 678
UMBC
February 7th, 2018
Announcements: Assignment 2

Due 11:59 AM, Monday 3/5

Topic: classification and neural models

Out: by Friday 2/9
Announcements: Course Project

Teams of 1-4

Three “checkins”:

Proposal: 3/14
Update: 4/9
Final submission: 5/23

Some novel aspect is needed

Ex 1: reimplement existing technique and apply to new domain
Ex 2: explore novel technique on existing problem
Ex 3: explore novel task

If you run into compute issues, talk to course staff. Options exist, e.g.,
https://aws.amazon.com/education/awseducate/
Recap from last time...
A Simple Linear Regression Model: From ERM

**loss function:** \( \ell = (y_i - \hat{y}_i)^2 \)

\[
\arg\min_w \sum_i (y_i - w^T x_i - b)^2 = \min_w (y - Xw - b)^T (y - Xw - b)
\]

least squares estimation
A Simple Linear Model for Classification

\[ y_i = w^T x_i \]

vector \( w \) of weights

label \( y_i \) (WLOG, binary \{0, 1\} value)

data point \( x_i \), as a vector of features

loss function:

\[ \ell = \begin{cases} 
1, & y_i w^T x_i < 0 \\
0, & y_i w^T x_i \geq 0 
\end{cases} = 1[y_i w^T x_i < 0] \]
Loss Function Example: 0-1 Loss

\[ \ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases} \]

**Problem:** not differentiable wrt \( \hat{y} \) (or \( \theta \))

**Solution 1:** is \( h(x) \) a conditional distribution \( p(y \mid x) \)? Use MAP

**Solution 2:** use a surrogate loss that approximates 0-1

**Solution 3:** is the data linearly separable? Perceptron can work
## Classification Evaluation: Accuracy, Precision, and Recall

### Accuracy: % of items correct

\[
\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + FP + FN + TN}}
\]

### Precision: % of selected items that are correct

\[
\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}
\]

### Recall: % of correct items that are selected

\[
\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}
\]

<table>
<thead>
<tr>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/Guessed</strong></td>
<td><strong>Not select/not guessed</strong></td>
</tr>
<tr>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>
Maximum Entropy (Log-linear) Models

\[ p(x \mid y) \propto \exp(\theta \cdot f(x, y)) \]

classify in one go
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

*Observed document*
Three people have been fatally shot and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.

We need to score the different combinations.
Score and Combine Our Possibilities

score₁(fatally shot, ATTACK)
score₂(seriously wounded, ATTACK)
score₃(Shining Path, ATTACK)
...
scoreₖ(department, ATTACK)
...

are all of these uncorrelated?

COMBINE

posterior probability of ATTACK
Score and Combine Our Possibilities

\[ \text{score}_1(\text{fatally shot, ATTACK}) \]
\[ \text{score}_2(\text{seriously wounded, ATTACK}) \]
\[ \text{score}_3(\text{Shining Path, ATTACK}) \]

... 

Q: What are the score and combine functions for Naïve Bayes?
Scoring Our Possibilities

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[
\text{score}(\text{, ATTACK}) = \text{score}_1(\text{fatally shot, ATTACK}) + \text{score}_2(\text{seriously wounded, ATTACK}) + \text{score}_3(\text{Shining Path, ATTACK}) + \ldots
\]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
What function... operates on any real number?

is never less than 0?
What function... 

operates on any real number?

is never less than 0?

\[ f(x) = \exp(x) \]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[
p(\text{ATTACK}) \propto \exp\left( \sum \text{score}_i(\text{fatally shot, ATTACK}) + \text{score}_2(\text{seriously wounded, ATTACK}) + \text{score}_3(\text{Shining Path, ATTACK}) + \ldots \right)
\]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

Learn the scores (but we’ll declare what combinations should be looked at)
Maxent Modeling

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[
p(\text{ATTACK}) \propto \exp \left( \sum \text{weight}_1 \cdot \text{applies}_1(\text{fatally shot, ATTACK}) + \text{weight}_2 \cdot \text{applies}_2(\text{seriously wounded, ATTACK}) + \text{weight}_3 \cdot \text{applies}_3(\text{Shining Path, ATTACK}) + \ldots \right)
\]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

Q: How do we define Z?
Normalization for Classification

\[ Z = \sum \exp \left( \begin{array}{c}
\text{weight}_1 \ast \text{applies}_1 \text{(fatally shot, } x) \\
\text{weight}_2 \ast \text{applies}_2 \text{(seriously wounded, } x) \\
\text{weight}_3 \ast \text{applies}_3 \text{(Shining Path, } x) \\
\text{...}
\end{array} \right) \]

\[ p(x \mid y) \propto \exp(\theta \cdot f(x, y)) \]

classify doc y with label x in one go
Q: What if none of our features apply?
Guiding Principle for Log-Linear Models

“[The log-linear estimate] is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information.”

Edwin T. Jaynes, 1957
Guiding Principle for Log-Linear Models

“[The log-linear estimate] is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information.”

Edwin T. Jaynes, 1957

\[ \exp(\theta \cdot f) \Rightarrow \exp(\theta \cdot 0) = 1 \]
Easier-to-write form

\[
\exp(\theta_1 f_1(\text{fatally shot, ATTACK}) + \theta_2 f_2(\text{seriously wounded, ATTACK}) + \theta_3 f_3(\text{Shining Path, ATTACK}) + \ldots)
\]
Easier-to-write form

\[ \exp(\theta_1 f_1(\text{fatally shot}, \text{ATTACK}) \theta_2 f_2(\text{seriously wounded}, \text{ATTACK}) \theta_3 f_3(\text{Shining Path}, \text{ATTACK}) \ldots) \]

\( K \) weights \( \rightarrow \) \( \text{features} \)
Easier-to-write form

\[ \exp(\theta \cdot f_{doc, ATTACK}) \]

- K-dimensional weight vector
- K-dimensional feature vector

dot product
Log-Linear Models

\[ p_\theta(x \mid y) \propto \exp(\theta^T f(x, y)) \]
Log-Linear Models

\[ p_\theta(x \mid y) \propto \exp(\theta^T f(x, y)) \]

- Feature function(s)
- Sufficient statistics
- "Strength" function(s)
Log-Linear Models

\[ p_\theta(x \mid y) \propto \exp(\theta^T f(x, y)) \]
Normalization for Classification

\[ Z = \sum_{\text{label } X} \exp\left( \text{weight}_1 \cdot \text{applies}_1(\text{fatally shot, } x) \right) + \exp\left( \text{weight}_2 \cdot \text{applies}_2(\text{seriously wounded, } x) \right) + \exp\left( \text{weight}_3 \cdot \text{applies}_3(\text{Shining Path, } x) \right) + \ldots \]

\[ p(x \mid y) \propto \exp(\theta \cdot f(x, y)) \]

classify doc y with label x in one go
Log-Linear Models

\[ p_{\theta}(x \mid y) = \frac{\exp(\theta^T f(x, y))}{\sum_{x'} \exp(\theta^T f(x', y))} \]
Connections to Other Techniques

Log-Linear Models
Connections to Other Techniques

Log-Linear Models

(Multinomial) logistic regression
Softmax regression
"Solution" 1: A Simple Probabilistic (Linear*) Classifier

Decision rule:

\[
\hat{y}_i = \begin{cases} 
0, & \sigma(w^T x_i + b) < 0.5 \\
1, & \sigma(w^T x_i + b) \geq 0.5 
\end{cases}
\]

Remember from "Linear regression"

Minimize posterior 0-1 loss:

\[
\min_w \sum_i \mathbb{E}_{\hat{y}_i} [1[y_ip(\hat{y}_i = 1|x_i) < 0]] = \\
\max_w \sum_i p(\hat{y}_i = y_i|x_i)
\]

Why MAP classifiers are reasonable

*linear not strictly required
Connections to Other Techniques

Log-Linear Models

(Multinomial) logistic regression
Softmax regression
Maximum Entropy models (MaxEnt)

as statistical regression
based in information theory
Connections to Other Techniques

Log-Linear Models

(Multinomial) logistic regression
Softmax regression
Maximum Entropy models (MaxEnt)
Generalized Linear Models

as statistical regression
based in information theory
a form of
Connections to Other Techniques

Log-Linear Models

- (Multinomial) logistic regression
- Softmax regression
- Maximum Entropy models (MaxEnt)
- Generalized Linear Models
- Discriminative Naïve Bayes

as statistical regression

based in information theory

a form of

viewed as
Connections to Other Techniques

Log-Linear Models

- (Multinomial) logistic regression
- Softmax regression
- Maximum Entropy models (MaxEnt)
- Generalized Linear Models
- Discriminative Naïve Bayes
- Very shallow (sigmoidal) neural nets

as statistical regression
based in information theory
a form of
viewed as
to be cool today :)
Objective = Full Likelihood?

\[ \prod_i p_\theta(x_i | y_i) \propto \prod_i \exp(\theta^T f(x_i, y_i)) \]

These values can have very small magnitude \(\Rightarrow\) underflow

Differentiating this product could be a pain
Log-Likelihood

\[ \log \prod_i p_{\theta}(x_i | y_i) = \sum_i \log p_{\theta}(x_i | y_i) = \sum_i \theta^T f(x_i, y_i) - \log Z(y_i) \]

Wide range of (negative) numbers
Sums are more stable

Differentiating this becomes nicer (even though Z depends on \( \theta \))
Log-Likelihood Gradient

Each component $k$ is the difference between:
Log-Likelihood Gradient

Each component $k$ is the difference between:

the total value of feature $f_k$ in the training data

$$
\sum_i f_k(x_i, y_i)
$$
Log-Likelihood Gradient

Each component $k$ is the difference between:

- the total value of feature $f_k$ in the training data
- the total value the current model $p_\theta$ thinks it computes for feature $f_k$

\[
\sum_i f_k(x_i, y_i) \quad \text{and} \quad \sum_i \mathbb{E}_p[f(x', y_i)]
\]

"Moment Matching"

A1 Q2 & A1 Q3, Eq-1 (what were the feature functions)?
Lesson 6
Log-Likelihood Gradient Derivation

\[ \nabla_\theta F(\theta) = \nabla_\theta \sum_i [\theta \cdot f(x_i, y_i) - \log Z(y_i)] \]
Log-Likelihood Gradient Derivation

\[ \nabla_{\theta} F(\theta) = \nabla_{\theta} \sum_i \left[ \theta \cdot f(x_i, y_i) - \log Z(y_i) \right] \]

\[ = \sum_i f(x_i, y_i) - \text{depends on } \theta \]

\[ Z(y_i) = \sum_{x'} \exp(\theta \cdot f(x', y_i)) \]
Log-Likelihood Gradient Derivation

\[
\nabla_\theta F(\theta) = \nabla_\theta \sum_i [\theta \cdot f(x_i, y_i) - \log Z(\ )) \\
= \sum_i f(x_i, y_i) - \sum_i \sum_{x'} \frac{\exp(\theta^T f(x', y_i))}{Z(y_i)} f(x', y_i)
\]

use the (calculus) chain rule

\[
\frac{\partial}{\partial \theta} \log g(h(\theta)) = \left( \frac{\partial g}{\partial h(\theta)} \right) \left( \frac{\partial h}{\partial \theta} \right)
\]

scalar \( p(y' \mid x_i) \)

vector of functions
Log-Likelihood Gradient Derivation

\[ \nabla_{\theta} F(\theta) = \nabla_{\theta} \sum_i \left[ \theta \cdot f(x_i, y_i) - \log Z(\ ) \right] \]

\[ = \sum_i f(x_i, y_i) - \sum_i \sum_{x'} \frac{\exp(\theta^T f(x', y_i))}{Z(y_i)} f(x', y_i) \]

use the (calculus) chain rule

\[ \frac{\partial}{\partial \theta} \log g(h(\theta)) = \left( \frac{\partial g}{\partial h(\theta)} \right) \left( \frac{\partial h}{\partial \theta} \right) \]

scalar \( p(x' \mid y_i) \)

vector of functions
Log-Likelihood Gradient Derivation

\[ \nabla_\theta F(\theta) = \nabla_\theta \sum_i [\theta \cdot f(x_i, y_i) - \log Z(\quad)] \]

\[ = \sum_i f(x_i, y_i) - \sum_i \sum_{x'} \frac{\exp(\theta^T f(x', y_i))}{Z(y_i)} f(x', y_i) \]

Do we want these to fully match?

What does it mean if they do?

What if we have missing values in our data?
Lesson 8---Regularization
Understanding Conditioning

\[ p(x \mid y) \propto \exp(\theta \cdot f(y)) \]

Is this a good posterior classifier? (no)
Connections to Other Techniques

Log-Linear Models

(Multinomial) logistic regression
Softmax regression
Maximum Entropy models (MaxEnt)
Generalized Linear Models
Discriminative Naïve Bayes

Very shallow (sigmoidal) neural nets

as statistical regression
based in information theory
a form of
viewed as
to be cool today :)