Lecture 4: Logistic Regression

Shuai Li
John Hopcroft Center, Shanghai Jiao Tong University
shuaili8.github.io

Last lecture

• Linear regression
  • Normal equation
  • Gradient methods
  • Examples
  • Probabilistic view
  • Applications
  • Regularization
Today’s lecture

• Discriminative / Generative Models
• Logistic regression (binary classification)
  • Cross entropy
  • Formulation, sigmoid function
  • Training—gradient descent
• More measures for binary classification (AUC, AUPR)
• Class imbalance
• Multi-class logistic regression
Discriminative / Generative Models
Discriminative / Generative Models

• Discriminative models
  • Modeling the dependence of unobserved variables on observed ones
  • also called conditional models.
  • Deterministic: \( y = f_\theta(x) \)
  • Probabilistic: \( p_\theta(y|x) \)

• Generative models
  • Modeling the joint probabilistic distribution of data
  • Given some hidden parameters or variables
    \[ p_\theta(x, y) \]
  • Then do the conditional inference
    \[ p_\theta(y|x) = \frac{p_\theta(x, y)}{p_\theta(x)} = \frac{p_\theta(x, y)}{\sum_{y'} p_\theta(x, y')} \]
Discriminative Models

• Discriminative models
  • Modeling the dependence of unobserved variables on observed ones
  • also called conditional models.
  • Deterministic: \( y = f_\theta(x) \)
  • Probabilistic: \( p_\theta(y|x) \)

• Directly model the dependence for label prediction
• Easy to define dependence on specific features and models
• Practically yielding higher prediction performance
• E.g. linear regression, logistic regression, k nearest neighbor, SVMs, (multi-layer) perceptrons, decision trees, random forest
Generative Models

• Generative models
  • Modeling the joint probabilistic distribution of data
  • Given some hidden parameters or variables
    \[ p_\theta(x, y) \]
  • Then do the conditional inference
    \[ p_\theta(y|x) = \frac{p_\theta(x, y)}{p_\theta(x)} = \frac{p_\theta(x, y)}{\sum_{y'} p_\theta(x, y')} \]

• Recover the data distribution [essence of data science]
• Benefit from hidden variables modeling
• E.g. Naive Bayes, Hidden Markov Model, Mixture Gaussian, Markov Random Fields, Latent Dirichlet Allocation
Discriminative Models vs Generative Models

• In General
  • A Discriminative model models the **decision boundary between the classes**
  • A Generative Model explicitly models the **actual distribution of each class**

• Example: Our training set is a bag of fruits. Only **apples** and **oranges**
  Each labeled. Imagine a post-it note stuck to the fruit
  • A generative model will model various attributes of fruits such as color, weight, shape, etc
  • A discriminative model might model color alone, **should that suffice** to distinguish apples from oranges
<table>
<thead>
<tr>
<th></th>
<th><strong>Discriminative model</strong></th>
<th><strong>Generative model</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>Directly estimate $P(y</td>
<td>x)$</td>
</tr>
<tr>
<td><strong>What's learned</strong></td>
<td>Decision boundary</td>
<td>Probability distributions of the data</td>
</tr>
<tr>
<td><strong>Illustration</strong></td>
<td><img src="image" alt="Discriminative model illustration" /></td>
<td><img src="image" alt="Generative model illustration" /></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>Regressions, SVMs</td>
<td>GDA, Naive Bayes</td>
</tr>
</tbody>
</table>
Linear Discriminative Models

• Discriminative model
  • modeling the dependence of unobserved variables on observed ones
  • also called conditional models
  • **Deterministic**: \( y = f_\theta(x) \)
  • Probabilistic: \( p_\theta(y|x) \)

• Linear regression model

\[
y = f_\theta(x) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j = \theta^\top x
\]

\[
x = (1, x_1, x_2, \ldots, x_d)
\]
Logistic Regression
From linear regression to logistic regression

• Logistic regression
  • Similar to linear regression
    • Given the numerical features of a sample, predict the numerical label value
    • E.g. given the size, weight, and thickness of the cell wall, predict the age of the cell
  • The values $y$ we now want to predict take on only a small number of discrete values
    • E.g. to predict the cell is benign or malignant
Example

- Given the data of cancer cells below, how to predict they are benign or malignant?
Logistics regression

- It is a Classification problem
  - Compared to regression problem, which predicts the labels from many numerical features

- Many applications
  - Spam Detection: Predicting if an email is Spam or not based on word frequencies
  - Credit Card Fraud: Predicting if a given credit card transaction is fraud or not based on their previous usage
  - Health: Predicting if a given mass of tissue is benign or malignant
  - Marketing: Predicting if a given user will buy an insurance product or not
Classification problem

• Given:
  • A description of an instance \(x \in X\)
  • A fixed set of categories: \(C = \{c_1, c_2, \ldots, c_m\}\)

• Determine:
  • The category of \(x : f(x) \in C\) where \(f(x)\) is a categorization function whose domain is \(X\) and whose range is \(C\)
  • If the category set binary, i.e. \(C = \{0, 1\}\) (\{false, true\}, \{negative, positive\}) then it is called binary classification
Binary classification

- Linearly separable
- Nonlinearly separable
Linear discriminative model

• Discriminative model
  • modeling the dependence of unobserved variables on observed ones
  • also called conditional models.
  • Deterministic: \( y = f_\theta(x) \)
  • Probabilistic: \( p_\theta(y|x) \)

• For binary classification
  • \( p_\theta(y = 1 | x) \)
  • \( p_\theta(y = 0 | x) = 1 - p_\theta(y = 1 | x) \)
Loss Functions
KL divergence

• Regression: mean squared error (MSE)
• Kullback-Leibler divergence (KL divergence)
  • Measure the dissimilarity of two probability distributions

\[
\begin{align*}
\mathbb{KL}(p||q) & \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \\
\mathbb{KL}(p||q) &= \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H}(p) + \mathbb{H}(p, q)
\end{align*}
\]

Question:
Which one is more similar to norm distribution?
KL divergence (cont.)

• Information inequality

\[ \mathbb{KL}(p||q) \geq 0 \] with equality iff \( p = q \).

• Entropy

  • \( \mathcal{H}(X) \triangleq - \sum_{k=1}^{K} p(X = k) \log_2 p(X = k) \)
  
  • Is a measure of the uncertainty
  
  • Discrete distribution with the maximum entropy is the uniform distribution

• Cross entropy

  • \( \mathcal{H}(p, q) \triangleq - \sum_k p_k \log q_k \)
  
  • Is the average number of bits needed to encode data coming from a source with distribution \( p \) when we use model \( q \) to define our codebook
Cross entropy loss

• Cross entropy
  • Discrete case: $H(p, q) = -\sum_x p(x) \log q(x)$
  • Continuous case: $H(p, q) = -\int_x p(x) \log q(x)$

• Cross entropy loss in classification:
  • Red line $p$: the ground truth label distribution.
  • Blue line $q$: the predicted label distribution.
Example for binary classification

• Cross entropy: $H(p, q) = -\sum_x p(x) \log q(x)$

• Given a data point $(x, 0)$ with prediction probability $q_\theta(y = 1|x) = 0.4$
  the cross entropy loss on this point is
  $L = -p(y = 0|x) \log q_\theta(y = 0|x) - p(y = 1|x) \log q_\theta(y = 1|x)$
  $= -\log(1 - 0.4) = \log \frac{3}{5}$

• What is the cross entropy loss for data point $(x, 1)$ with prediction probability
  $q_\theta(y = 1|x) = 0.3$
Cross entropy loss for binary classification

• Loss function for data point \((x, y)\) with prediction model 
  \(p_\theta(\cdot | x)\)

  is

  \[
  L(y, x, p_\theta) \\
  = -1_{y=1} \log p_\theta(1| x) - 1_{y=0} \log p_\theta(0| x) \\
  = -y \log p_\theta(1| x) - (1 - y) \log (1 - p_\theta(1| x))
  \]
Cross entropy loss for multiple classification

• Loss function for data point \((x, y)\) with prediction model 
\[ p_\theta(\cdot | x) \]
is
\[
L(y, x, p_\theta) = - \sum_{i=1}^{m} 1_{y=C_k} \log p_\theta(C_k | x)
\]
Binary Classification
Binary classification: linear and logistic
Binary classification: linear and logistic

• Linear regression:
  • Target is predicted by $h_\theta(x) = \theta^T x$

• Logistic regression
  • Target is predicted by $h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$
  where
  $$\sigma(z) = \frac{1}{1 + e^{-z}}$$
  is the logistic function or the sigmoid function
Properties for the sigmoid function

• \( \sigma(z) = \frac{1}{1+e^{-z}} \)
  • Bounded in \((0,1)\)
  • \( \sigma(z) \to 1 \) when \( z \to \infty \)
  • \( \sigma(z) \to 0 \) when \( z \to -\infty \)

• \( \sigma'(z) = \frac{d}{dz} \frac{1}{1+e^{-z}} = -(1 + e^{-z})^{-2} \cdot (-e^{-z}) \)
  \[ = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} \]
  \[ = \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}}\right) \]
  \[ = \sigma(z)(1 - \sigma(z)) \]
Logistic regression

• Binary classification

\[ p_\theta(y = 1|x) = \sigma(\theta^\top x) = \frac{1}{1 + e^{-\theta^\top x}} \]
\[ p_\theta(y = 0|x) = \frac{e^{-\theta^\top x}}{1 + e^{-\theta^\top x}} \]

• Cross entropy loss function

\[ \mathcal{L}(y, x, p_\theta) = -y \log \sigma(\theta^\top x) - (1 - y) \log(1 - \sigma(\theta^\top x)) \]

• Gradient

\[ \frac{\partial \mathcal{L}(y, x, p_\theta)}{\partial \theta} = -y \frac{1}{\sigma(\theta^\top x)} \sigma(z)(1 - \sigma(z))x - (1 - y) \frac{-1}{1 - \sigma(\theta^\top x)} \sigma(z)(1 - \sigma(z))x \]

\[ = (\sigma(\theta^\top x) - y)x \]
\[ \theta \leftarrow \theta + \eta(y - \sigma(\theta^\top x))x \]

\[ \frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z)) \]

\[ \theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \]
Label decision

• Logistic regression provides the probability

\[
p_\theta(y = 1|x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}
\]

\[
p_\theta(y = 0|x) = \frac{e^{-\theta^T x}}{1 + e^{-\theta^T x}}
\]

• The final label of an instance is decided by setting a threshold \( h \)

\[
\hat{y} = \begin{cases} 
1, & p_\theta(y = 1|x) > h \\
0, & \text{otherwise}
\end{cases}
\]
How to choose the threshold

• Precision-recall trade-off
  • Precision = \frac{TP}{TP+FP}
  • Recall = \frac{TP}{TP+FN}

• Higher threshold
  • More FN and less FP
  • Higher precision
  • Lower recall

• Lower threshold
  • More FP and less FN
  • Lower precision
  • Higher recall

\hat{y} = \begin{cases} 
1, & p_\theta(y = 1|x) > h \\
0, & \text{otherwise} 
\end{cases}
Example

• We have the heights and weights of a group of students
  • Height: in inches,
  • Weight: in pounds
  • Male: 1, female, 0

• Please build a Logistic regression model to predict their genders
Example (cont.)

• As there are only two features, height and weight, the logistic regression equation is:
  \[ h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}} \]

• Solve it by gradient descent

• The solution is \( \theta = \begin{bmatrix} 0.69254 \\ -0.49269 \\ 0.19834 \end{bmatrix} \)

There will be a lab hw on logistic regression
• Threshold $h = 0.5$

• Decision boundary is $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

• Above the decision boundary lie most of the blue points that correspond to the Male class, and below it all the pink points that correspond to the Female class.

• The predictions won’t be perfect and can be improved by including more features (beyond weight and height), and by potentially using a different decision boundary (e.g. nonlinear)
Example 2

- A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

<table>
<thead>
<tr>
<th>Hours</th>
<th>Pass</th>
<th>Hours</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0</td>
<td>2.75</td>
<td>1</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>3.00</td>
<td>0</td>
</tr>
<tr>
<td>1.00</td>
<td>0</td>
<td>3.25</td>
<td>1</td>
</tr>
<tr>
<td>1.25</td>
<td>0</td>
<td>3.50</td>
<td>0</td>
</tr>
<tr>
<td>1.50</td>
<td>0</td>
<td>4.00</td>
<td>1</td>
</tr>
<tr>
<td>1.75</td>
<td>0</td>
<td>4.25</td>
<td>1</td>
</tr>
<tr>
<td>1.75</td>
<td>1</td>
<td>4.50</td>
<td>1</td>
</tr>
<tr>
<td>2.00</td>
<td>0</td>
<td>4.75</td>
<td>1</td>
</tr>
<tr>
<td>2.25</td>
<td>1</td>
<td>5.00</td>
<td>1</td>
</tr>
<tr>
<td>2.50</td>
<td>0</td>
<td>5.50</td>
<td>1</td>
</tr>
</tbody>
</table>
Example 2 (cont.)

\[ h_\theta(x) = \frac{1}{1 + e^{-(1.5046 \times \text{hours} - 4.0777)}} \]
Interpretation of logistic regression

• Given a probability $p$, the odds of $p$ is defined as $odds = \frac{p}{1-p}$

• The logit is defined as the log of the odds: $\ln(odds) = \ln\left(\frac{p}{1-p}\right)$

• Let $\ln(odds) = \theta^T x$, we will have $\ln\left(\frac{p}{1-p}\right) = \theta^T x$, and

$$p = \frac{1}{1 + e^{-\theta^T x}}$$

• So in logistic regression, the logit of an event (predicted positive)’s probability is defined as a result of linear regression
More Measures for Classification
Confusion matrix

- Remember what we have learned about the confusion matrix

- **Precision**: the ratio of true class 1 cases in those with prediction 1
  \[
  \text{Prec} = \frac{TP}{TP + FP}
  \]

- **Recall**: the ratio of cases with prediction 1 in all true class 1 cases
  \[
  \text{Rec} = \frac{TP}{TP + FN}
  \]

\[
F1 = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}
\]

- These are the basic metrics to measure the classifier
Area Under ROC Curve (AUC)

- A performance measurement for classification problem at various thresholds settings
- Tells how much the model is capable of distinguishing between classes
- The higher, the better
- Receiver Operating Characteristic (ROC) Curve
  - TPR against FPR
  - TPR/Recall/Sensitivity = $\frac{TP}{FP}$
  - Sensitivity = $S_{TP+FN}$
  - FPR = 1 - Specificity = $\frac{TN+FP}{TP+FN}$
AUC (cont.)

- It’s the relationship between TPR and FPR when the threshold is changed from 0 to 1.
- In the top right corner, threshold is 0, and everything is predicted to be positive, so both TPR and FPR is 1.
- In the bottom left corner, threshold is 1, and everything is predicted to be negative, so both TPR and FPR is 0.
- The size of the area under this curve (AUC) is an important metric to binary classifier.
- Perfect classifier get AUC=1 and random classifier get AUC = 0.5.

TPR: true positive rate
FPR: false positive rate
AUC (cont.)

- It considers all possible thresholds.
- Various thresholds result in different true/false positive rates.
- As you decrease the threshold, you get more true positives, but also more false positives.
- From a random classifier you can expect as many true positives as false positives. That’s the dashed line on the plot. AUC score for the case is 0.5. A score for a perfect classifier would be 1. Most often you get something in between.
AUC example

AUC = 0.75

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>1</td>
</tr>
<tr>
<td>0.85</td>
<td>0</td>
</tr>
<tr>
<td>0.77</td>
<td>1</td>
</tr>
<tr>
<td>0.72</td>
<td>1</td>
</tr>
<tr>
<td>0.61</td>
<td>0</td>
</tr>
<tr>
<td>0.48</td>
<td>1</td>
</tr>
<tr>
<td>0.42</td>
<td>0</td>
</tr>
<tr>
<td>0.33</td>
<td>0</td>
</tr>
</tbody>
</table>
Precision recall curve

• The precision recall curve, or pr curve, is another plot to measure the performance of binary classifier.

- It’s the relationship between Precision and Recall when the threshold is changed from 0 to 1
- It’s more complex than the ROC curve
- The size of the area under this curve is an important metric to binary classifier
- It can handle imbalanced dataset
- Usually, the classifiers gets lower AUPR value than AUC value
AUPR examples

- **Perfect**: 1
- **Good**: 0.92
- **Random**: 0.56
Class Imbalance
Class imbalance

• Down sampling
  • Sample less on frequent class

• Up sampling
  • Sample more on infrequent class

• Hybrid Sampling
  • Combine them two
Weighted loss functions

\[ L(y, x, p_\theta) = -y \log p_\theta(1|x) - (1 - y)\log (1 - p_\theta(1|x)) \]

\[ L(y, x, p_\theta) = -w_1 y \log p_\theta(1|x) - w_0 (1 - y)\log (1 - p_\theta(1|x)) \]
Multi-Class Logistic Regression
Multi-class classification

\[ L(y, x, p_\theta) = - \sum_{i=1}^{m} 1_{y=c_k} \log p_\theta(C_k|x) \]
Multi-Class Logistic Regression

• Class set \( C = \{c_1, c_2, \ldots, c_m\} \)

• Predicting the probability of \( p_\theta(y = c_j | x) \)

\[
p_\theta(y = c_j | x) = \frac{e^{\theta_j^T x}}{\sum_{k=1}^{m} e^{\theta_k^T x}} \quad \text{for } j = 1, \ldots, m
\]

• Softmax
  • Parameters \( \theta = \{\theta_1, \theta_2, \ldots, \theta_m\} \)
  • Can be normalized with \( m-1 \) groups of parameters
Multi-Class Logistic Regression

• Learning on one instance \((x, y = c_j)\)
  • Maximize log-likelihood

\[
\max_{\theta} \log p_\theta(y = c_j | x)
\]

• Gradient

\[
\frac{\partial \log p_\theta(y = c_j | x)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \log \frac{e^{\theta_j^T x}}{\sum_{k=1}^{m} e^{\theta_k^T x}}
= x - \frac{\partial}{\partial \theta_j} \log \sum_{k=1}^{m} e^{\theta_k^T x}
= x - \frac{e^{\theta_j^T x} x}{\sum_{k=1}^{m} e^{\theta_k^T x}}
\]
Summary

• Discriminative / Generative Models
• Logistic regression (binary classification)
  • Cross entropy
  • Formulation, sigmoid function
  • Training—gradient descent
• More measures for binary classification (AUC, AUPR)
• Class imbalance
• Multi-class logistic regression
Next Lecture

SVM
Questions?