Barbora Hladká  
hladka@ufal.mff.cuni.cz

Martin Holub  
holub@ufal.mff.cuni.cz

Charles University,  
Faculty of Mathematics and Physics,  
Institute of Formal and Applied Linguistics
Outline

• Logistic regression

• Evaluation of binary classifiers
A task of binary classification: $Y = \{0, 1\}$

Decision boundary takes a form of function $f$ and partitions a feature space into two sets, one for each class.
Hyperplane is a linear decision boundary of the form

$$\Theta^T x = 0$$

where direction of $$\langle \theta_1, \theta_2, \ldots, \theta_m \rangle$$ is perpendicular to the hyperplane and $$\theta_0$$ determines position of the hyperplane with respect to the origin.
• point if $m = 1$, line if $m = 2$, plane if $m = 3$, …
• we can use hyperplane for classification so that

$$f(x) = \begin{cases} 
1 & \text{if } \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m \geq 0 \\
0 & \text{if } \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m < 0 
\end{cases}$$

• linear classifiers classify examples using hyperplanes
Binary classification
Can we use linear regression?
Binary classification
Can we use linear regression?

Fit the data with a linear function $f$

$$f(x) = \Phi^T x$$
Binary classification
Can we use linear regression?

Classify
- if $f(x) \geq 0.5$, predict 1
- if $f(x) < 0.5$, predict 0

$f(x) = \Phi^T x$
Binary classification
Can we use linear regression?

Add one more training instance

\[ f(x) = \Phi^T x \]

(Yes) 1

0.5

(No) 0
Binary classification
Can we use linear regression?

We are heading for the logistic regression algorithm.
Logistic regression is a classification algorithm.

Its target hypothesis $f$ for a binary classification has a form of \textbf{sigmoid function}

$$f(x; \Theta) = \frac{1}{1 + e^{-\Theta^\top x}} = \frac{e^{\Theta^\top x}}{1 + e^{\Theta^\top x}}$$

\begin{itemize}
  \item $g(z) = \frac{1}{1 + e^{-z}}$
  \item $\lim_{z \to +\infty} g(z) = 1$
  \item $\lim_{z \to -\infty} g(z) = 0$
\end{itemize}
Logistic regression

\[ f(x; \Theta) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1}} \]
Logistic regression

\[ f(x; \Theta) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1}} \]
Logistic regression
Classification rule

Predict a target value using $f(\mathbf{x}; \hat{\Theta})$ so that

- if $f(\mathbf{x}; \hat{\Theta}) \geq 0.5$, i.e. $\hat{\Theta}^\top \mathbf{x} \geq 0$, predict 1
- if $f(\mathbf{x}; \hat{\Theta}) < 0.5$, i.e. $\hat{\Theta}^\top \mathbf{x} < 0$, predict 0
Interpretation of $f(x; \Theta)$: it models the conditional probability $Pr(y = 1|x; \Theta)$

$$f(x; \Theta) = Pr(y = 1|x; \Theta)$$

1. categorical attribute $Y = \{0, 1\}$

2. $y = \theta_0 + \theta_1 x_1 \cdots + \theta_m x_m$, see above $\rightarrow$ model $Pr(Y = y|x)$, e.g. $Pr(Y = 1|x)$

3. $Pr(Y = 1|x) = \theta_0 + \theta_1 x_1 \cdots + \theta_m x_m$, see above

4. Model odds $(Pr(Y = 1|x)) = \frac{Pr(Y=1|x)}{Pr(Y=0|x)} = \frac{Pr(Y=1|x)}{1-Pr(Y=1|x)} \in (0, +\infty)$
Odds, odds ratio

odds = \Pr(\text{success})/\Pr(\text{failure})

Titanic example

```r
> d <- read.csv("train.csv")
> attach(d)
> table(Sex, Survived)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>81</td>
<td>233</td>
</tr>
<tr>
<td>male</td>
<td>468</td>
<td>109</td>
</tr>
</tbody>
</table>

> detach()
```

• the odds of surviving for male:
  \( \Pr(\text{Survived} = 1|\text{Sex} = \text{male})/\Pr(\text{Survived} = 0|\text{Sex} = \text{male}) = \frac{109}{486} = 0.23 \)

• the odds of surviving for female:
  \( \Pr(\text{Survived} = 1|\text{Sex} = \text{female})/\Pr(\text{Survived} = 0|\text{Sex} = \text{female}) = \frac{233}{81} = 2.88 \)

• the ratio of the odds for female to the odds for male \( 2.88/0.23 = 12.52 \)
5. Transform \((0, +\infty)\) to \((-\infty, +\infty)\): model

\[
\logit(\Pr(Y = 1|\mathbf{x})) = \ln(\text{odds}(\Pr(Y = 1|\mathbf{x}))) = \ln\left(\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})}\right)
\]

6. Use linear regression

\[
\ln\left(\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})}\right) = \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m
\]

i.e.,

\[
\Pr(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1 - \cdots - \theta_m x_m}}
\]

\[
f(\mathbf{x}_i; \Theta) = \Pr(Y_i = 1|\mathbf{x}_i; \Theta) = \frac{1}{1 + e^{-\Theta^\top \mathbf{x}_i}}
\]
Parameter interpretation

Binary features

• Use female = \{1, 0\} instead of Sex = \{female, male\}

• in Linear regression \( y = \theta_0 + \theta_1 \times \text{female} \)
  • \( \theta_0 \) is the average \( y \) for male
  • \( \theta_0 + \theta_1 \) is the average \( y \) for female
  • \( \theta_1 \) is the average difference in \( y \) between female and male

• in Logistic regression \( p = \Pr(Survive = 1|\mathbf{x}, \Theta) \), \( \ln \frac{p}{1-p} = \theta_0 + \theta_1 \times \text{female} \)
  • If female == 0
    • \( p = p_1 \rightarrow \ln\left(\frac{p_1}{1-p_1}\right) = \theta_0 \rightarrow \frac{p_1}{1-p_1} = e^{\theta_0} \)
    • the intercept \( \theta_0 \) is the log odds for men
  • If female == 1
    • \( p = p_2 \rightarrow \frac{p_2}{1-p_2} = e^{\theta_0+\theta_1} \)
    • odds ratio = \( \frac{p_2}{1-p_2} / \frac{p_1}{1-p_1} = e^{\theta_1} \)
    • the parameter \( \theta_1 \) is the log odds ratio between female and male
Parameter interpretation
Numerical features

• $\theta_i$ gives an average change in $\logit(f(x))$ with one-unit change in $A_i$ holding all other features fixed
Parameter estimates

- **Loss function**

\[
L(\Theta) = - \sum_{i=1}^{n} y_i \log P(y_i|x_i; \Theta) + (1 - y_i) \log (1 - P(y_i|x_i; \Theta))
\]

See Maximum Likelihood Principle for derivation of this loss function.

- **Optimization problem**

\[
\Theta^* = \arg\min_{\Theta} L(\Theta)
\]
Parameter estimates

\[ L(\Theta) = -\sum_{i=1}^{n} y_i \log P(y_i|x_i; \Theta) + (1 - y_i) \log(1 - P(y_i|x_i; \Theta)) \]
Parameter estimates

$L(\Theta) = -\sum_{i=1}^{n} y_i \log P(y_i|x_i; \Theta) + (1 - y_i) \log (1 - P(y_i|x_i; \Theta))$
repeat until convergence {

Θ^{K+1} := Θ^K - α \nabla f(Θ^K)

}

- α is a positive step-size hyperparameter

l.e. simultaneously update \( \theta_j \), \( j = 1, \ldots, m \)

\[
\theta_j^{K+1} := \theta_j^K - \alpha \frac{1}{n} \sum_{i=1}^{n} (f(x_i; \Theta^K) - y_i)x_{ij}
\]
Non-linear decision boundary
Non-linear decision boundary

- For hyperplane: \( f(x) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m) \)
- Let \( f(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2) \) (a higher degree polynomial)
- Assume \( \theta_0 = -1, \theta_1 = 0, \theta_2 = 0, \theta_3 = 1, \theta_4 = 1 \)
- Predict \( y = 1 \) if \(-1 + x_1^2 + x_2^2 \geq 0\), i.e. \( x_1^2 + x_2^2 \geq 1 \)
Logistic regression

Summary

Classification of $x$ by $\hat{f}^*$

1. Project $x$ onto $\hat{\Theta}^*$ to convert it into a real number $z$ in the range $(-\infty, +\infty)$
   - i.e. $z = \hat{\Theta}^* \top x$

2. Map $z$ to the range $\langle 0, 1 \rangle$ using the sigmoid function $g(z) = 1/(1 + e^{-z})$

3. Classify $x$ using a classification rule
Multi-class classification

$|Y| = N, \ N \geq 3$

- **One-to-all**
  - train $N$ binary classifiers $f_k$ for the pair $k$-th class and $\{1, \cdots, N\} \setminus \{k\}$ classes
  - classify $x$ into the class $k^* = \operatorname{argmax}_k f_k(x)$

- **One-to-one**
  - train $\binom{N}{2}$ binary classifiers $f_i$ for each pair of classes
  - classify $x$ into the class $k^* = \max_{k=1,\ldots,N} \sum_{i=1}^{\binom{N}{2}} \delta(f_i(x) = k)$
Logistic regression
Multi-class classification

One-to-all
**Confusion matrix** is a square matrix indexed by all possible target class values.

Task: Assign the correct sense of the word *line* in a sentence.

```
** Comparing the predicted values with the true senses **

<table>
<thead>
<tr>
<th>Prediction</th>
<th>cord</th>
<th>division</th>
<th>formation</th>
<th>phone</th>
<th>product</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cord</td>
<td>268</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>division</td>
<td>3</td>
<td>280</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>formation</td>
<td>13</td>
<td>3</td>
<td>225</td>
<td>4</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>phone</td>
<td>25</td>
<td>5</td>
<td>2</td>
<td>293</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>product</td>
<td>51</td>
<td>10</td>
<td>39</td>
<td>32</td>
<td>1442</td>
<td>72</td>
</tr>
<tr>
<td>text</td>
<td>12</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>28</td>
<td>262</td>
</tr>
</tbody>
</table>
```

Correctly predicted examples are displayed on the diagonal.
In binary classification tasks examples are sometimes regarded as divided into two disjoint subsets:

- **positive examples** – “to be retrieved” (ones)
- **negative examples** – “not to be retrieved” (zeros)

### Example confusion matrix for binary classification

```r
> table(test.true, test.pred)

prediction
0 1
true 0 580 69
      1 37 144
```
Evaluation of binary classifiers

Confusion matrix

<table>
<thead>
<tr>
<th>True class</th>
<th>Predicted class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td></td>
<td>(TP)</td>
<td>(FN)</td>
</tr>
<tr>
<td>Negative</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td>(FP)</td>
<td>(TN)</td>
</tr>
</tbody>
</table>

Explanation

- ‘Trues’ are examples correctly classified
- ‘Falses’ are examples incorrectly classified
- ‘Positives’ were predicted as positives (correctly or incorrectly)
- ‘Negatives’ were predicted as negatives (correctly or incorrectly)
Proportion of correctly predicted test examples

- Known positives
- Predicted positives
- False negatives
- True positives
- False positives
### Basic performance measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>$\frac{TP}{TP+FP}$</td>
</tr>
<tr>
<td>Recall/Sensitivity</td>
<td>$\frac{TP}{TP+FN}$</td>
</tr>
<tr>
<td>Specificity</td>
<td>$\frac{TN}{TN+FP}$</td>
</tr>
<tr>
<td>1-Specificity (FPR)</td>
<td>$\frac{FP}{TN+FP}$</td>
</tr>
<tr>
<td>Accuracy</td>
<td>$\frac{(TP+TN)}{(TP+FP+TN+FN)}$</td>
</tr>
</tbody>
</table>

Very often you need to **combine both good precision and good recall**. Then you usually use **balanced F-score**, so called **F-measure**

$$F = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
Sensitivity vs. Specificity

Perfect classifier – no error

Reality – errors
Sensitivity vs. Specificity

100% sensitive classifier

100% specific classifier
Sensitivity vs. specificity

<table>
<thead>
<tr>
<th>Test Value</th>
<th>Number of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Disease</td>
<td>Without Disease</td>
</tr>
<tr>
<td>100% sensitivity</td>
<td>100% specificity</td>
</tr>
</tbody>
</table>

Threshold

True Negative | False Negative | False Positive | True Positive

NPFL054, 2023 Hladká & Holub Lecture 6, page 37/43
An **ROC curve** plots True Positive Rate vs. False Positive Rate at different classification thresholds (see p. 6).
Evaluation of binary classifiers

AUC measure

**Area Under ROC** (\(=\) AUC)
is a measure of how good is a distinguishing property of classifier
Evaluation of binary classifiers

ROC & AUC
Evaluation of binary classifiers

ROC & AUC
Summary of Examination Requirements

- Decision boundary, hyperplane, classification rule
- Logistic regression, sigmoid function, probabilistic formulation
- Multi-class classification
- Confusion matrix for binary classification
- Basic performance measures, ROC curve, AUC
For more details refer to

- Ng Andrew, Machine Learning online at Stanford
  (https://class.coursera.org/ml-2012-002/class/index)

  (http://www-stat.stanford.edu/~tibs/ElemStatLearn/), Section 4.4