STA 314: Statistical Methods for Machine Learning I

Lecture 7 - Logistic regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- In classification, $X \in \mathcal{X}$ and $Y \in C = \{0, 1, \ldots, K-1\}$.

- The Bayes rule

$$\arg\max_{k \in C} \mathbb{P}\{Y = k \mid X = \mathbf{x}\}, \qquad \forall \mathbf{x} \in \mathcal{X}$$

  has the smallest expected error rate.

- For binary classification, our goal is to estimate

$$p(\mathbf{x}) = \mathbb{P}\{Y = 1 \mid X = \mathbf{x}\}, \qquad \forall \mathbf{x} \in \mathcal{X}.$$

Logistic Regression

Logistic Regression is a parametric approach that assumes parametric structure on

$$p(\mathbf{x}) = \mathbb{P}(Y = 1 \mid X = \mathbf{x}).$$

- It assumes

$$p(\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}.$$

  The function $f(t) = e^t/(1 + e^t)$ is called the logistic function. $\beta_0, \ldots, \beta_p$ are the parameters.

- It is easy to see that we always have $0 \leq p(\mathbf{x}) \leq 1$.

- Note that $p(\mathbf{x})$ is **NOT** a linear function either in $\mathbf{x}$ or in $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.

Logistic Regression

- A bit of rearrangement gives

$$\underbrace{\frac{p(\mathbf{x})}{1-p(\mathbf{x})}}_{\text{odds}} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p},$$

$$\underbrace{\log\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right]}_{\text{log-odds (a.k.a. logit)}} = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p.$$

odds $\in [0,\infty)$ and log-odds $\in (-\infty,\infty)$.

- Similar interpretation as linear models.

- How to estimate $\boldsymbol{\beta}$?

Maximum Likelihood Estimator (MLE)

Given $\mathcal{D}^{train} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ with $y_i \in \{0, 1\}$, we estimate the parameters by **maximizing the likelihood** of $\mathcal{D}^{train}$.

The maximum likelihood principle

We seek the estimates of parameters such that the fitted probability are the closest to the individual's observed outcome.

Cont'd: MLE under logistic regression

Recipe of computing the MLE:

1. Write down the likelihood, as always!
2. Solve the optimization (maximization) problem.

The MLE has many nice properties!

- Asymp. consistent
- Asymp. normal
- Asymp. efficient

Inference under logistic regression

Let $\hat{\boldsymbol{\beta}}$ be the MLE.

- Z-statistic is similar to t-statistic in regression, and is defined as

$$\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}, \qquad \forall j \in \{0, 1, \dots, p\}.$$

- It produces p-value for testing the null hypothesis

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0.$$

  A large (absolute) value of the z-statistic or small p-value indicates evidence against H_0.

Example: Default data

Consider the Default data using balance, income, and student status as predictors.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

Prediction at **different levels** under logistic regression

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ be the MLE.

- Prediction of **the logit** at $\mathbf{x} \in \mathcal{X}$:

$$\widehat{\text{logit}}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

- Prediction of **the conditional probability** $\mathbb{P}(Y = 1 \mid X = \mathbf{x})$:

$$\hat{\mathbb{P}}(Y = 1 \mid X = \mathbf{x}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}$$

- Prediction of **the label** Y (i.e. *classification*) at $X = \mathbf{x}$:

$$\hat{y} = \begin{cases} 1, & \text{if} \quad \hat{\mathbb{P}}(Y = 1 \mid X = \mathbf{x}) \geq 0.5; \\ 0, & \text{otherwise.} \end{cases}$$

Prediction of $\mathbb{P}(Y = 1 \mid X = \mathbf{x})$

Consider the Default data with student status as the only feature.

What is the probability of default for a student?

To fit the model, we encode student status as 1 for student and 0 otherwise.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Metrics used for evaluating classifiers

In classification, we have several metrics that can be used to evaluate a given classifier.

- The most commonly used metric is the overall classification accuracy.
- For binary classification, there are a few more out there.....

Logistic Regression on the Default Data

- Classify whether or not an individual will default on the basis of credit card balance and student status.
- **The confusion matrix** of fitted logistic regression

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Type of Errors for binary classification

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

1. The training error rate is $(23 + 252)/10000 = 2.75\%$.
2. **False positive rate (FPR)**: The fraction of negative examples that are classified as positive: $23/9667 = 0.2\%$ in default data.
3. **False negative rate (FNR)**: The fraction of positive examples that are classified as negative: $252/333 = 75.7\%$ in default data.[1]

[1] For a credit card company that is trying to identify high-risk individuals, the error rate 75.7% among individuals who default is unacceptable.

Types of Errors for binary classification

Q: How to modify the logistic classifier to lower the FNR?

- The current classifier is based on the rule

$$\hat{\mathbb{P}}(\text{default} = \textit{yes} \mid X = \mathbf{x}) \geq 0.5.$$

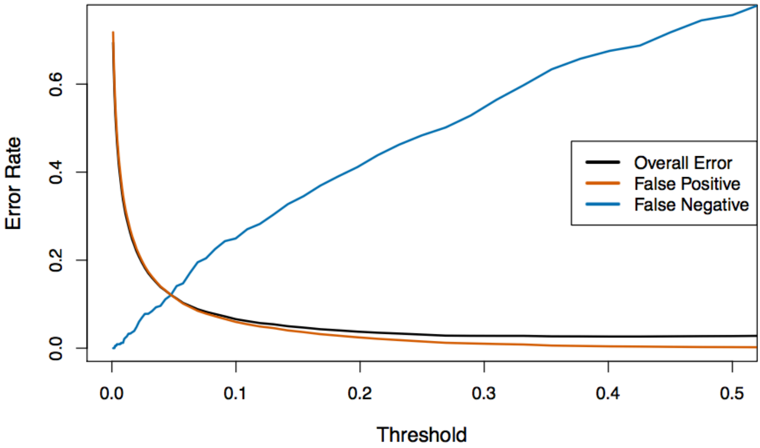
- To lower FNR, we reduce the number of negative predictions. Classify $X = \mathbf{x}$ to *yes* if

$$\hat{\mathbb{P}}(Y = \textit{yes} \mid X = \mathbf{x}) \geq t.$$

for some $t < 0.5$.

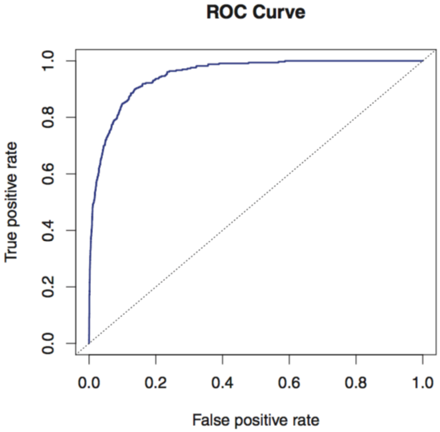
Trade-off between FPR and FNR

We can achieve better balance between FPR and FNR by varying the threshold:



ROC Curve

The **ROC curve** is a popular graphic for simultaneously displaying FPR and TPR = 1 - FNR for all possible thresholds.



The overall performance of a classifier, summarized over all thresholds, is given by the area under the curve (**AUC**). High AUC is good.

More metrics in the binary classification

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

The above also defines **sensitivity** and **specificity**.

Computation of the MLE under Logistic Regression

General steps of computing the MLE:

- Write down the likelihood, as always!
- Solve the optimization problem.

Likelihood under Logistic Regression

For simplicity, let us set $\beta_0 = 0$ such that

$$p(\mathbf{x}) = \frac{e^{\mathbf{x}^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^\top \boldsymbol{\beta}}}, \qquad 1 - p(\mathbf{x}) = \frac{1}{1 + e^{\mathbf{x}^\top \boldsymbol{\beta}}}.$$

The data consists of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ with

$$y_i \sim \text{Bernoulli}(p(\mathbf{x}_i)), \qquad p(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}, \quad 1 \le i \le n.$$

- What is the likelihood of y_i?

Likelihood under Logistic Regression

The likelihood of each data point $(\mathbf{x}_i, y_i)$ at any $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = [p(\mathbf{x}_i)]^{y_i} [1 - p(\mathbf{x}_i)]^{1-y_i}$$

with

$$p(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}.$$

The joint likelihood of all data points is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [p(\mathbf{x}_i)]^{y_i} [1 - p(\mathbf{x}_i)]^{1-y_i}.$$

Log-likelihood under Logistic Regression

The log-likelihood at any $\boldsymbol{\beta}$ is

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log\left\{\prod_{i=1}^{n}\left[p(\mathbf{x}_i)\right]^{y_i}\left[1-p(\mathbf{x}_i)\right]^{1-y_i}\right\} \\
&= \sum_{i=1}^{n}\left[y_i\log(p(\mathbf{x}_i)) + (1-y_i)\log(1-p(\mathbf{x}_i))\right] \\
&= \sum_{i=1}^{n}\left[y_i\log\left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right) + \log(1-p(\mathbf{x}_i))\right] \\
&= \sum_{i=1}^{n}\left[y_i\mathbf{x}_i^\top\boldsymbol{\beta} - \log\left(1+e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right)\right].
\end{aligned}$$

How to compute the MLE?

How do we maximize the log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log\left(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right) \right]$$

for logistic regression?

- It is equivalent to minimize $-\ell(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$.
- No direct solution: taking derivatives of $\ell(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ and setting them to 0 doesn't have an explicit solution.
- Need to use iterative procedure.