Classification

Wenda Zhou

June 15, 2017

Classification

Classification is a common problem in prediction.

1. Identifying credit card fraud
2. Medical diagnostic
3. Predicting device failure

Linear models for classification

Natural idea for classification: use logistic regression.

$$\operatorname{logit} p = \alpha + \beta x \tag{1}$$

Linear models for classification

Natural idea for classification: use logistic regression.

$$\operatorname{logit} p = \alpha + \beta x \tag{1}$$

Can predict p by using

$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}x}}{1 + e^{\hat{\alpha}+\hat{\beta}x}} \tag{2}$$

Linear models for classification

Natural idea for classification: use logistic regression.

$$\operatorname{logit} p = \alpha + \beta x \tag{1}$$

Can predict p by using

$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}x}}{1 + e^{\hat{\alpha}+\hat{\beta}x}} \tag{2}$$

However, need to go from $\hat{p}$ to 0-1 answer.

Precision vs recall

Need to choose cut-off for $\hat{p}$.

Precision vs recall

Need to choose cut-off for $\hat{p}$.
Have a precision-recall tradeoff. Can visualize this in a precision-recall curve.

Precision-recall tradeoff

Lower threshold will yield to higher recall. Higher threshold to higher precision.

Linear discriminant analysis

Bayes inspired method. Let $Y = 0, 1$ be the label of each observation, and let X be the covariate (continuous).
We will suppose that:

$$\begin{cases} X \mid Y = 0 \sim \mathcal{N}(\mu_1, \sigma^2) \\ X \mid Y = 1 \sim \mathcal{N}(\mu_2, \sigma^2) \end{cases} \tag{3}$$

We will also place a prior on Y given by $\mathrm{P}(Y = k) = \pi_k$.

Linear discriminant analysis

Bayes inspired method. Let $Y = 0, 1$ be the label of each observation, and let X be the covariate (continuous).
We will suppose that:

$$\begin{cases} X \mid Y = 0 \sim \mathcal{N}(\mu_1, \sigma^2) \\ X \mid Y = 1 \sim \mathcal{N}(\mu_2, \sigma^2) \end{cases} \tag{3}$$

We will also place a prior on Y given by $\mathrm{P}(Y = k) = \pi_k$.
Predict class according to posterior of Y, given by:

$$\mathrm{P}(Y = k \mid X) \propto f_X(x \mid y)\pi_k \tag{4}$$

Linear discriminant analysis

Thus predict the class for which the discriminant δ_k is largest, where

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \tag{5}$$

Linear discriminant analysis

Thus predict the class for which the discriminant δ_k is largest, where

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \tag{5}$$

Use plug-in estimators for μ_k and σ^2.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{6}$$

Naive Bayes

Another Bayes inspired method, but for $X_1, \ldots, X_n$ discrete (usually binary).

$$\begin{aligned}\mathrm{P}(Y = k \mid X_1 = x_1, \ldots, X_n = x_n) \propto \\ \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n \mid Y = k)\, \mathrm{P}(Y = k)\end{aligned} \tag{7}$$

Select class with largest probability.

Naive Bayes

$P(X_1 = x_1, \ldots, X_n = x_n \mid Y = k)$ difficult to estimate. Make conditional independence assumption that:

$$P(X_1 = x1, \ldots, X_n = x_n \mid Y = k) = \\ P(X_1 = x_1 \mid Y = k)\, P(X_2 = x_2 \mid Y = k) \cdots P(X_n = x_n \mid Y = k) \tag{8}$$

Naive Bayes

$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n \mid Y = k)$ difficult to estimate. Make conditional independence assumption that:

$$\mathrm{P}(X_1 = x1, \ldots, X_n = x_n \mid Y = k) = \\ \mathrm{P}(X_1 = x_1 \mid Y = k)\,\mathrm{P}(X_2 = x_2 \mid Y = k) \cdots \mathrm{P}(X_n = x_n \mid Y = k) \tag{8}$$

Can now estimate from the data

$$\mathrm{P}(X_1 = x_1 \mid Y = k) = \frac{\#\text{ of times with } X_1 = x_1 \text{ for class } k}{\#\text{ occurence of class } k} \tag{9}$$

Naive Bayes

$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n \mid Y = k)$ difficult to estimate. Make conditional independence assumption that:

$$\mathrm{P}(X_1 = x1, \ldots, X_n = x_n \mid Y = k) = \\ \mathrm{P}(X_1 = x_1 \mid Y = k)\,\mathrm{P}(X_2 = x_2 \mid Y = k) \cdots \mathrm{P}(X_n = x_n \mid Y = k) \tag{8}$$

Can now estimate from the data

$$\mathrm{P}(X_1 = x_1 \mid Y = k) = \frac{\#\text{ of times with } X_1 = x_1 \text{ for class } k}{\#\text{ occurence of class } k} \tag{9}$$

Adequate even for very large number of X_i.

Naive Bayes

Example: text data

Naive Bayes very successful in textual data (e.g. spam classification).

Naive Bayes

Example: text data

Naive Bayes very successful in textual data (e.g. spam classification).

Text data

Not easy to operate on: need to transform to some numerical covariates.

Bag of words

Simplest representation of text data: bag of words.
Consider a document. For each word i in dictionary, have variable x_i with $x_i = 1$ if word is in document, and $x_i = 0$ otherwise.

Bag of words

Simplest representation of text data: bag of words.
Consider a document. For each word i in dictionary, have variable x_i with $x_i = 1$ if word is in document, and $x_i = 0$ otherwise.

$$\text{document} = (x_1, x_2, \ldots, x_{100000}) \tag{10}$$

Naive Bayes

Example: spam

1. Obtain training dataset: spam email + non-spam email
2. For each word in dictionary, compute Naive Bayes estimate

$$\mathrm{P}(\text{business} \mid \text{spam}) = \frac{\text{number of spam emails with word business}}{\text{total number of spam emails}} \tag{11}$$