STATS 769

Classification II

Yong Wang
Department of Statistics
The University of Auckland

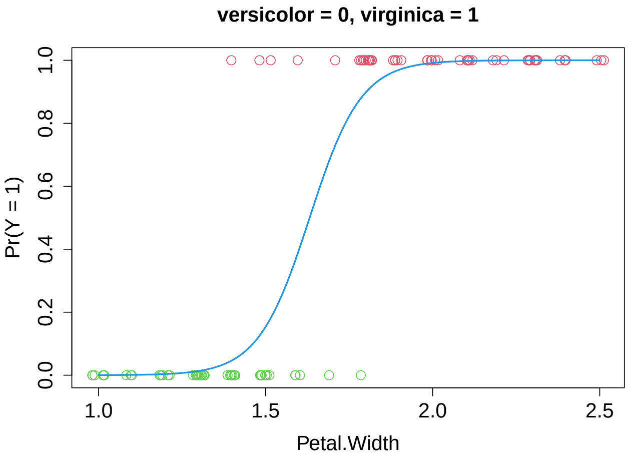
2021

Basic Classification Methods

- Linear discriminant analysis
- Quadratic discriminant analysis
- Naive Bayes

- Logistic regression
- Generalised additive models
- K-nearest neighbours

Iris Data



(`Petal.Width` has some jittering)

Simple Logistic Regression

- For the simple logistic model here, we have

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

- The estimates $(\widehat{\beta}_0, \widehat{\beta}_1)$ can be found by maximum likelihood.
- We should then classify an observation x as class 1 if

$$\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x),$$

  or class 0 if otherwise.
- We can denote the log-odds by

$$\begin{aligned}\delta_{10}(x) &= \log[\mathbb{P}(Y = 1|X = x)] - \log[\mathbb{P}(Y = 0|X = x)] \\ &= \beta_0 + \beta_1 x.\end{aligned}$$

- These are already posterior probabilities, so we don't need π_0 or π_1 as in LDA/QDA.
- The decision boundary is the point x where $\delta_{10}(x) = 0$.
  - It is where $\mathbb{P}(Y = 1|X = x) = \frac{1}{2}$ or $\beta_0 + \beta_1 x = 0$.

Multiple Logistic Regression

- Similarly,

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots\beta_p x_p}}.$$

- The estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ can be found by maximum likelihood.
- The rest is pretty much the same as for simple linear regression.
- Again, let

$$\begin{aligned}\delta_{10}(x) &= \log[\mathbb{P}(Y = 1|X = x)] - \log[\mathbb{P}(Y = 0|X = x)] \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.\end{aligned}$$

- The decision boundary is determined by $\delta_{10}(x) = 0$, or $\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p = 0$, which defines a hyperplane in $\mathbb{R}^p$.

Multiple Logistic Regression II

- Just like multiple linear regression, one can also carry out subset selection or use regularisation methods.

Multinomial Logistic Regression

- The multiple logistic regression can be easily expended to cope with more than two classes.
- Assume there are $J > 2$ classes.
- We first choose one baseline class (as class 0).
- Then for class $j = 1, \ldots, J-1$, we let

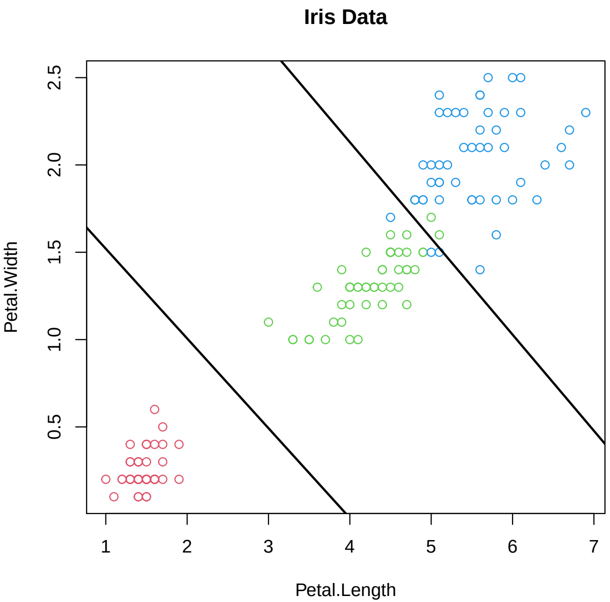
$$\mathbb{P}(Y = j|X = x) = \frac{e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p}}{1+\sum_{l=1}^{J-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

- Then

$$\mathbb{P}(Y = 0|X = x) = 1 - \sum_{j=1}^{J-1} \mathbb{P}(Y = j|X = x).$$

- Choose class j^*, if $\mathbb{P}(Y = j^*|X = x)$ is the largest out of all $j = 0, \ldots, J-1$.

Decision Boundaries of Multiple Logistic Regression



Multinomial Logistic Regression II

- Note that, for $j = 1, \dots, J-1$,

$$\begin{aligned}\delta_{j0}(x) &= \log\left[\frac{\mathbb{P}(Y=j|X=x)}{\mathbb{P}(Y=0|X=x)}\right]\\ &= \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p.\end{aligned}$$

- Hence, for $j, k \in \{1, \dots, J-1\}$,

$$\begin{aligned}\delta_{jk}(x) &= \log\left[\frac{\mathbb{P}(Y=j|X=x)}{\mathbb{P}(Y=k|X=x)}\right]\\ &= (\beta_{j0} - \beta_{k0}) + (\beta_{j1} - \beta_{k1})x_1 + \cdots + (\beta_{jp} - \beta_{kp})x_p.\end{aligned}$$

- Function $\delta_{jk}(x)$ obtained directly for a pair of classes is not exactly the same as $\delta_{jk}(x)$ obtained by fitting a multinomial logistic regression.

Logistic Regression: Discussion

- Using conditional probabilities, hence there is no need to make distributional assumptions for X which hardly hold in practice.
- Safer and robust
- Has some numerical issues if $\mathbb{P}(Y = j|X = x)$ becomes 1 or 0 (numerically).

Generalised Additive Models

- Using GAMs for classification is very similar to their use for regression.
- Let's take multiple logistic regression as an example:

$$\begin{aligned}\delta_{10}(x) &= \log[\mathbb{P}(Y=1|X=x)] - \log[\mathbb{P}(Y=0|X=x)] \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.\end{aligned}$$

- We can just replace each linear term $\beta_j x_j$ with a smoothing spline.

K-nearest Neighbours

- For any target point x', find its K-nearest neighbours, in Euclidean distance (most likely):

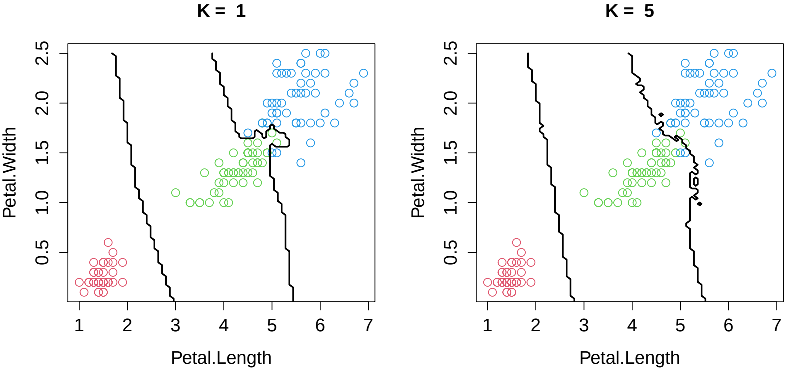
$$d(x, x') = \sqrt{(x_1 - x_1')^2 + \cdots + (x_p - x_p')^2}.$$

- Use the majority rule for classification: Choose the class with the most votes from the K neighbours.
- This is the same as choosing the largest value of

$$\mathbb{P}(Y = j | X = x') = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x')} I(y_i = j), \quad j = 1, \ldots, J,$$

  where $\mathcal{N}_K(x')$ denotes the set of K-nearest neighbours of x'.
- This is a memory-based method.
- It can also be used for regression problems (by taking the mean of the response).

Decision Boundaries of KNN



We can choose a good value for K, by using a data resampling technique.

Recommended Readings

ISLv2 (basics):

- Sections 2.2.3, 4.3, 7.7.2
- Labs: Sections 4.7.2, 4.7.5, 4.7.6

ESL (advanced):

- Sections 4.4, 9.1, 13.3