THE UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS

STAT 447B Methods for Statistical Learning (2014/15 Term 1)
Lab 3

1/2 October, 2014

1 Review

Logistic Regression

- If our response Y is binary, we may want to use **logistic regression**. For $i = 1, \ldots, n$, we assume $Y_i|\boldsymbol{X}_i \sim Bin(1, p_i)$ where

$$p_i = \mathbb{E}(Y_i|\boldsymbol{X}_i) = \mathbb{P}(Y_i = 1|\boldsymbol{X}_i) = \frac{\exp\{\boldsymbol{X}_i'\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{X}_i'\boldsymbol{\beta}\}}$$

  with $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ being the (column) vector of covariates for observation i and parameters respectively. Parameter estimation is through maximum likelihood, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log\left[p_i^{y_i}(1-p_i)^{1-y_i}\right] = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i)\right].$$

  We call this a **generalized linear model** because of the linear component $\boldsymbol{X}_i'\boldsymbol{\beta}$. Logistic regression can be performed using the function `glm` in R, specifying the family as `binomial`.

- Similar to ordinary regression, we can make our fit more flexible by changing $\boldsymbol{X}_i'\boldsymbol{\beta}$ to a polynomial or spline basis. The function `gam` can be used for this purpose.

Classification

- A **classifier** is a function that admits covariates/features as input and a class as output. The objective of classification is to predict which group a future observation belongs to given the features associated with it. The training data are thus used to "train" the classifier. For example, in fitting a logistic regression model, a possible classification rule is to predict a future observation i as 1 if $\hat{p}_i \geq 0.5$ and 0 otherwise.

- The **apparent error rate** measures the proportion of misclassification among the training data. It is obvious that this is not a good measure of a model's predictive ability.

- Under the **0-1 loss function**, the classification rule to minimize the expected loss is to select the category at which the probability of occurrence is the highest given the covariates, i.e. category j should be chosen if $\mathbb{P}(G = j|\boldsymbol{X}) \geq \mathbb{P}(G = i|\boldsymbol{X})$ for all $i \neq j$. We can thus utilize the estimated probabilities $\hat{\mathbb{P}}(G = i|\boldsymbol{X})$ to build our classifier. Sometimes this quantity comes from the Bayes' rule:

$$\mathbb{P}(G = i|\boldsymbol{X}) = \frac{f_{\boldsymbol{X}|G}(\boldsymbol{x}|G = i)\mathbb{P}(G = i)}{f_{\boldsymbol{X}}(x)} \propto f_{\boldsymbol{X}|G}(\boldsymbol{x}|G = i)\mathbb{P}(G = i). \tag{1}$$

  In parametric estimation we thus specify the distribution of $\boldsymbol{X}$ conditional on each group as well as the overall probability of being in each group.

- **Linear discriminant analysis** (LDA) assumes each group has normally distributed features with common variance in applying (1). It is important to note that the boundaries of the classifiers in both the logistic and LDA models are hyperplanes of the covariate space. For instance, if there are two covariates, then the separators are necessarily straight lines.

2 Data analysis

In this lab we focus on classification based on the very famous *Iris* flower data set. It was originally used by Sir Ronald Fisher to illustrate discriminant analysis. For simplicity we focus on two of the four features in the original data set, and two out of the three possible classes (species). The objective is to use the flowers' sepal and petal lengths (in cm) to predict which species they belong to. In our data set (available on the lab webpage), $Y = 0$ and 1 represent the species *Iris versicolor* and *Iris virginica* respectively.

2.1 Visualize the data

Q1 There are two covariates and one response (with two classes). Plot petal length versus sepal length to display the data (with both axes spanning from 3 to 8 using the attributes `xlim` and `ylim`), and remember to use different colours to represent the two species.

Hint: The `col` attribute in `plot` accepts a vector as its value, corresponding to the colour at each plotted point.

2.2 Logistic classifier vs LDA

We will compare the sensitivity of the two classifiers to data perturbation.

Q2 First focus on the original data and find the boundary of the logistic classifier.

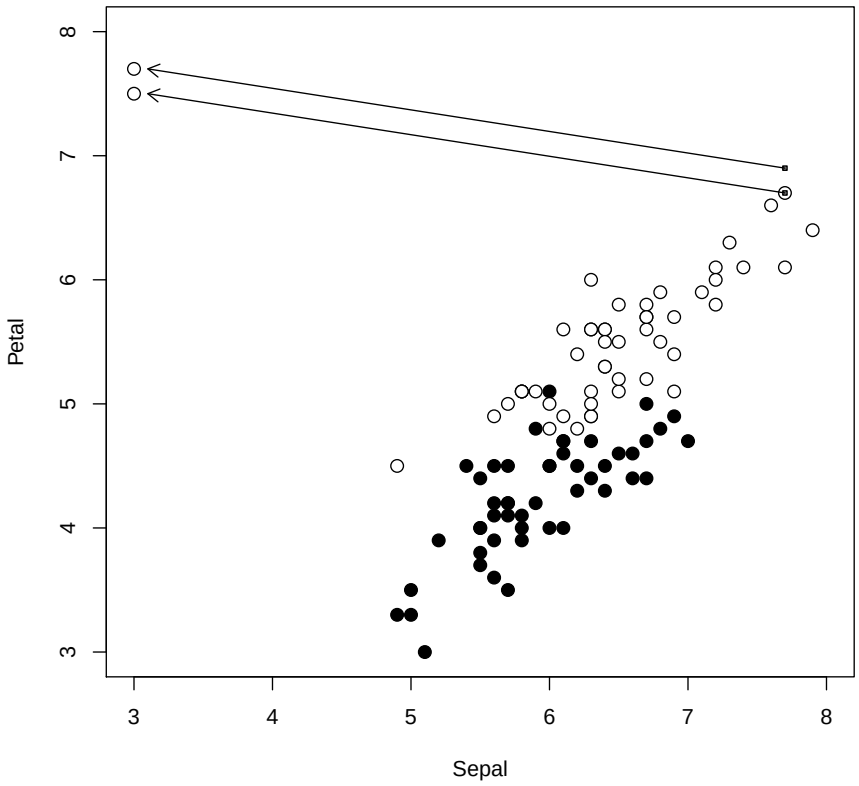
1. Using `glm`, perform logistic regression on the data set with both sepal and petal lengths as covariates.
2. Create a grid of values that covers the range of the covariates. Predict on this grid and plot the boundary (i.e. the collection of points with predicted value 0.5).

   Hint: In the previous lab, you learned the trick of creating a sequence of x values and predict on them, so that the predicted curve can be plotted. You are basically doing the same thing here, albeit more involved as there are *two* covariates – you will need to create a grid instead of a sequence. Some notes:

   - Use the `expand.grid` function to create a matrix holding the grid of values resulting from two vectors. Refer to the code for the lecture on September 30 if necessary.
   - In using `predict`, you need to make sure that your matrix of new data has the same column names as the variables used in fitting the model. Use `colnames` to access and change them if needed. You will also need to use `type='response'`. (Why?)
   - To plot the boundary, you can use the `contour` function. The first two parameters should contain the vectors you put into `expand.grid`, while the third should be a matrix of predicted probabilities, with dimensions the same as the two vectors. You will also need to instruct `contour` to plot *just* the 0.5 contour.

Q3 Repeat **Q2** using LDA. Pay attention to how you extract the predicted probabilities using `predict`. Compare the two boundaries you obtained.

Q4 Now we change the values of the covariates for the 68th and 69th observations. The effect of this change is illustrated in the figure on the next page.



The modified data set is also available on the lab webpage. Repeat **Q2** and **Q3** using this data. What happens to the boundaries of the classifiers?

Q5 Which classifier appears more "logical" with regard to this particular modification of the data? Can you explain why the two classifiers behave differently?