Statistical Learning, Tutorato #2

Veronica Vinciotti, Marco Chierici

March 15, 2021

Exercise 1

A number of common distributions belong to the Exponential Dispersion Family. We have seen the Normal and Binomial distribution in class (which give rise to linear and logistic regression, respectively). In this exercise, you will explore other common distributions:

- Poisson distribution (for count response variables):

$$f(y) = e^{-\lambda}\frac{\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

- Gamma distribution (for non-negative, typically skewed, continuous response variables):

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-y\frac{\nu}{\mu}}, \qquad y > 0$$

In each case:

- Show that the distribution is a member of the EDF (i.e. identify the functions $a(\phi)$, $b(\theta)$ and $c(y, \phi)$)
- Calculate the mean and variance functions using $b(\theta)$. Notice how in both cases the variance depends on the mean (ie heteroskedasticity naturally embedded in the model)
- Identify the canonical link

Exercise 2

Consider a a linear discriminant analysis with one predictor ($p = 1$) and balanced classes ($\pi_1 = \pi_2 = 0.5$). Use the derivation from lectures to show that the decision boundary in this simple case is given by $x = (\mu_1 + \mu_2)/2$.

Exercise 3

Consider the `Default` data set already analyzed in Tutorato #1.

- Plot some relevant numerical or graphical summaries of the data, showing the discriminative power of the predictors.
- Split the data randomly into train/test.
- Perform a logistic regression model on the training data to predict `default` and calculate the test error.
- Perform LDA on the training data to predict `default` and calculate the test error of this model.
- Perform QDA on the training data to predict `default` and calculate the test error of this model.
- You will notice that the error rates are all rather small. Is this a sign of very good models? Reflect on this by: 1) Calculating the performance of a null model that predicts always to the non-default class 2) Calculating specificity and sensitivity of the models from the confusion matrix. What do you conclude?
- Given that the class variable is very unbalanced, it is more informative to compare the methods based on the ROC curve. Plot the ROC curves of each model, computing also the areas under the curves (AUCs). Based on the results, which method would you use for future predictions?

Hints

- LDA and QDA are implemented in the functions `lda()`, `qda()` within the package `MASS`. Their syntax is the same as that of `glm` or `lm`.