Materials Informatics - Fall 2017

Computer Project 1 – Solutions

Due on: Oct 10

Assignment 1: In this part you will simulate two QSPRs X_1 and X_2 and a property Y, using a Gaussian model, where the prior probabilities are $P(Y=0) = P(Y=1) = \frac{1}{2}$, the means are at the points $\mu_0 = (0,0)$ and $\mu_1 = (1,1)$, and common covariance matrix

$$\Sigma = \sigma^2 \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right)$$

where $0 \le \rho \le 1$ is the correlation coefficient between the predictors X_1 and X_2 . Note that this is a homoskedastic Gaussian model (so that the optimal decision boundary is a hyperplane).

(a) Show that the optimal classification error for this model is given by:

$$\varepsilon^* = \Phi\left(-\frac{1}{\sqrt{2}\sigma\sqrt{1+\rho}}\right),$$

where $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-x^2/2} dx$ is the CDF of a standard normal random variable.

- (b) Plot on the same axes the optimal error for $\rho = 0$ (independent predictors), $\rho = 0.4$, $\rho = 0.6$, $\rho = 0.8$, $\rho = 1$ (perfect correlation, $X_1 = aX_2 + b$) as a function of σ . Describe the effect of ρ and σ on classification difficulty.
- (c) For $\rho=0.2$, $\sigma=1$, simulate a sample of size n=20 from the model. Obtain the LDA decision boundary corresponding to these data. Plot the data (using O's for class 0 and X's for class 1), with the superimposed decision boundary for the optimal classifier and the designed LDA classifiers. Describe what you see.
- (d) Compute the error of the LDA classifier by two methods: (1) using the formula given in class; (2) simulating a test sample of size m = 500 from the Gaussian model and forming a test-set error estimate. Repeat this for n = 40, n = 60, n = 80, and n = 100 (there is no need to plot the classifiers). Plot the classification error curves as a function of n (one curve for each method of computation). Describe what you see.

Assignment 2: Here we are going to use a real material data set to do a simple classification experiment. The Excel file containing the data is available on e-Campus. The data come from the publication

T. Yonezawa, K. Suzuki, S. Ooki and A. Hashimoto, "The Effect of Chemical Composition and Heat Treatment Conditions on Stacking Fault Energy for Fe-Cr-Ni Austenitic Stainless Steel." *Metall and Mat Trans A* (2013) 44: 5884.

The predictors are the element content in the metalic material (Columns A-Q), while the property to be predicted is the stacking fault energy (SFE, Column T). We will categorize the SFE into two classes, low (SFE ≤ 35) and high (SFE ≥ 45). We are therefore throwing out the middle values.

- (a) Pre-processing: discard all predictors that do not have at least 60% nonzero values; from the data that remain, remove the rows (samples) that contain any zero values; randomly split the remaining data into training (20%) and test data (80%). Reject any training sample that contains more than 55% from one of the populations (repeat the sampling).
- (b) Using Welch's two-sample t-test on the *training data*, produce a table with the predictors, T statistic, and p-value, ordered with largest absolute T statistics at the top.
- (c) Pick the top two predictors and design an LDA classifier. Plot the data with the superimposed LDA classifier. Estimate the classification error using the test set.
- (d) Repeat for the top three, four, and five predictors. Estimate the errors on the testing data (there is no need to plot the classifiers). What can you observe?