BM 59D Machine Learning

Homework 2

Due: November 7, 2017

1) a) Find the class priors, $P(C_i)$, class likelihoods (assume Gaussian), $p(x|C_i)$, and the evidence, $p(x)$, using the training data provided.

   Find the posteriors, $P(C_i|x)$, risks (0/1 loss, no rejection), and the discriminants $g_i(x) = logP(x|C_i) + logP(C_i)$.

   Plot all suitably, and notice the class boundaries with respect to all. Are they the same? Different? Why? What happens if I assume equal priors?

   Then analytically solve for $g_1(x) = g_2(x)$. Does the plot verify your solution?

   For minimum risk (loss), report the decision rule and the confusion matrices (one for the training set and one for the validation set)

   b) Assuming that the positive class is C_1 and the negative class is C_2, let $\lambda_{12} = 0.5$. Note that this is equivalent to assuming a loss matrix of

$$\lambda = \begin{bmatrix} 0 & 0.5 \\ 1 & 0 \end{bmatrix}$$

   What happens to the class boundaries (commenting based on the plot is sufficient) and to the confusion matrices in this case? Comment on your results.

   c) In addition to the asymmetric loss introduced above, consider an extra action of rejection with a loss of 0.2. Now, what happens to the class boundaries (commenting based on the plot is sufficient) and to the confusion matrices (you may add an extra column for rejection)? Comment on your results.

   d) Assuming that this classifier is used in clinical diagnosis, what are the TN, TP, FN, FP, sensitivity, specificity, PPV, NPV, and accuracy (total correct classification) rates, for the first case above (1a)? Comment on asymmetric loss and rejection in this context.

2) Perform least squares regression, by fitting polynomials of order 0 to 9, using the data provided. For each, find the associated error value. For each, also find the error between the fitted polynomial and the actual curve ($r = x^3 - x + 1$). Plot both errors with respect to the polynomial order. Do this for $\sigma_n = 0.5$ (first column in the data), $\sigma_n = 0.3$ (second column in the data) and $\sigma_n = 0.1$ (third column in the data), where σ_n is the standard deviation of the additive (and zero-mean Gaussian) noise. Comment on your results.

Please use MATLAB. Please do not use the built-in or community-shared functions to calculate the probabilities, or, to perform regression.

Please send your report by e-mail (To: ipekbounee@gmail.com, Sbj: BM59D HW2), and give your codes in the appendix at the end of your report so that I can trace back if unexpected results appear. I prefer one single document including all the material rather than a folder including a report together with several m-files.