STAT 538 Homework 1
Out January 5, 2011
Due January 12, 2011

©Marina Meilă
mmp@stat.washington.edu

Submit the code used to solve these problems through the `Assignments` web page. Turn in the required solutions (without the code) in class on the due date. Only the paper part of the homework is graded.

Problem 1 – Classifiers in 1 dimension

This homework will make use of the (one-dimensional) data set $\mathcal{D}$ contained in the file `hw1-1d-train.dat`. The file contains one example `x y` per row, like this

```
-2.028238 -1
-4.819767 -1
-4.081050 -1
```

... Use this data set to answer the questions below.

a. Estimating a generative classifier Assume the distributions $g_\pm(x)$ are normal distributions $N(\mu_\pm, 1)$. Estimate $\mu_\pm$ and $p = P(Y = 1)$ from the data.

Show that the generative classifier $f_g(x)$ corresponding to the generative distributions above is of the form

$$f_g(x) = \begin{cases} +1 & \text{if } x > \theta_g \\ -1 & \text{if } x < \theta_g \\ 0 & \text{if } x = \theta_g \end{cases} \tag{1}$$

for some real number θ_g. What is the expression of θ_f? Estimate its numerical value.

b. Estimating a nearest neigbor (NN) classifier Find the labels of the points $x = 0, 1, 2, -0.1$ by NN[1] classification using $\mathcal{D}$.

Plot the decision regions of the NN classifier determined by $\mathcal{D}$, i.e. plot the function $f_{NN}(x) \in \{\pm 1\}$ versus x.

[**Extra credit**] Explain how this function can be computed efficiently (i.e in $\mathcal{O}(N)$ operations and one sweep through the data).

c. Estimating a Linear classifier Show that for $x \in \mathbb{R}$ any linear classifier

[1] More precisely by 1-NN classification. Optionally: try 3-NN, 5-NN.

is of the form

$$f_L(x) \;=\; \mathrm{sgn}(sx - \theta_L) \tag{2}$$

with $s = \pm 1$ and $\theta_L \in \mathbb{R}$. Find the s and the θ_L that minimize the misclassification error L_{01} on the data set $\mathcal{D}$.

d. The Bayes loss The data were generated from two normal distributions with means $\mu_\pm = \pm 2$ and $p = 1/3$. Use this true data distribution to answer the following questions.

Calculate $P(Y = 1|x)$ as a function of x, μ_+, μ_-, p. Show that it has the form $1/(1 + e^{ax-b})$. Then give the expression for $P(Y = -1|x)$.

Suppose x is fixed; how should $y(x)$ be chosen so that the probability of making an error for this x is minimized? Plot $P(error|x)$ vs. x; mark the locations of $\mu_\pm$ on the graph.

From this, derive the expression for f_*, the Bayes-optimal classifier. *For question d I expect a (short) mathematical derivation that ends in a simple, explicit formula for a, b and f_*. The f_* you will obtain depends on a parameter θ_*; calculate its numerical value.*

Calculate the value of the Bayes loss L_{01}^*.

e. Make a plot of $pg_+(x)$ and $(1-p)g_-(x)$ on the same graph. Mark also the locations of $\mu_\pm, \theta_g, \theta_L, \theta_*$ on the graph.