Problem 1 – How is the K-nearest neighbor classifier affected by sampling noise?

Assume that we have a binary classification problem where \( x \in \mathbb{R}^2 \) and \( P_{XY} = P_Y P_{X|Y} \). \( P_Y(+1) = 0.5 \), \( P_{X|Y=\pm 1} = \text{Normal}(\mu_{\pm}, I_2) \) with \( I_2 \) the unit matrix of order 2 and \( \mu_{\pm} = [\pm 1.6 \ 0]^T \).

In this problem we will study by simulation how the decisions of the K-NN classifier fluctuate when the training set is resampled. Repeat questions \( a, b, c, d \) for \( K = 1, 3, 7, 11, 15, 19, \ldots 40 \) and optionally for other values of \( K \).

a. Generate simulation data (you aren’t required to show anything for this question, nor for \( b, c, d \))

1. Sample a test set \( \tilde{D} \) of size \( \tilde{N} = 1000 \) or larger from \( P_{XY} \)
2. Implement the K-NN classifier.
   Repeat for \( b = 1 \) to \( B \) with \( B \geq 30 \)
   (a) Sample a data set \( D_b \) of size \( N = 100 \) from \( P_{XY} \)
   (b) Denote by \( f_b \) the K-NN classifier based on \( D_b \). Calculate \( \tilde{y}^{ib} = f_b(\tilde{x}^i) \) for \( \tilde{x}^i \in \tilde{D} \) (The predictions of \( f_b \) on test sample).
   (c) Calculate \( \tilde{l}_b = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} 1[\tilde{y}^{ib} \neq \tilde{y}^i] \) for \( (x^i, y^i) \in D_b \). (How well does \( f_b \) fit the training set)
   (d) Calculate \( L_b \) the (estimated) expected loss of \( f_b \)
   \[
   L_b = L(f_b) = \frac{1}{N} \sum_{(\tilde{x}^i, \tilde{y}^i) \in \tilde{D}} 1[f_b(\tilde{x}^i) \neq \tilde{y}^i]
   \] (1)

b. Calculate the average and variance of the expected losses; denote \( L = \text{average}(L_b) \).
   This is a Monte Carlo estimate of the expected loss of the K-NN on this problem, when the sample size is \( N = 100 \).
c. For each point $i$ in the test set, calculate

$$p_i = \frac{\sum_{b=1}^{B}(\hat{y}_{ib} + 1)/2}{B}.$$  

(2)

This is the (empirical) probability that point $\tilde{x}^i$ is labeled $+$. Then calculate the (empirical) variance of the labeling of $i$, i.e. the averaged variance of $f(\tilde{x}^i)$.

$$V = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} p_i(1 - p_i)$$  

(3)

d. Calculate $\bar{l}$ the mean of $\hat{l}_b$.

e. Show how the above statistics depend on $K$. For the values of $K$ you used, plot $L, \bar{l}, V$ versus $K$ on the same graph. For $L$ and $\bar{l}$ also show error bars equal to $\text{stdev}(L_b)$, $\text{stdev}(\bar{l}_b)$ respectively.

f. Interpret the graphs in e. Which graphs informs about the variance of $f$, the K-NN classifier? What does it show about the influence of $K$ on the classifier variance?

g. Which graph informs about the bias of $f$, the K-NN classifier? What does it show about the influence of $K$ on the classifier bias?

j. Give a formula or algorithm for calculating/estimating the Bayes error $L^*$ for this problem. Assume that you have all the information in the first paragraph, and a computer to run simulations.

Calculate the actual value of $L^*$ using your method. (Optionally, plot it as a horizontal line on the graph in question e.)

Problem 2 – Classifiers in 1 dimension

This homework will make use of the (one-dimensional) data set $D$ contained in the file `hw2-1d-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.

For this problem and in general: if a result is already in the lecture notes you can use it as is. No need to derive it again. In particular in b below, specialize the formula from Lecture 1 to this case. In a, only numerical results required.
a. Assume the distributions $g_{\pm}(x) = P_{X|Y=\pm 1}(x)$ are normal distributions $N(\mu_{\pm}, 1)$. Estimate $\mu_{\pm}$ and $p = P(Y = 1)$ from the data.

b. **Estimating a generative classifier (LDA)** Denote by $f_{g}(x)$ the LDA classifier for this problem. Write $f_{g}$ in the form below

$$f_{g}(x) = \begin{cases} +1 & \text{if } x > \theta_{g} \\ -1 & \text{if } x < \theta_{g} \\ 0 & \text{if } x = \theta_{g} \end{cases}, \quad (4)$$

find the expression of $\theta_{g}$ as a function of $\mu_{\pm}, p$ and evaluate its numerical value from the estimates you obtained in a.

c. **Estimating a Linear classifier** Show that for $x \in \mathbb{R}$ any linear classifier is of the form

$$f_{L}(x) = \text{sgn}(sx - \theta_{L}) \quad (5)$$

with $s = \pm 1$ and $\theta_{L} \in \mathbb{R}$.  

Plot the value of the empirical classification error $\hat{l}_{01}$ on $D$ as a function of $\theta_{L}$ for $s = 1$.

Then find the $s$ and the $\theta_{L}$ that minimize the $\hat{l}_{01}$ on the data set $D$.

d. **The Bayes loss** The data were generated from two normal distributions with means $\mu_{+} = 2, \mu_{-} = -1.2$ and $p = 1/3$. Use this true data distribution to answer the following questions. 

Calculate $P(Y = 1|x)$ as a function of $x$ and the true $\mu_{\pm}, p$. You know from Lecture 1 that $P(Y = 1|x)$ has the form $1/(1 + e^{ax-b})$. Find the numerical values of $a$ and $b$.

Then, write the expression of the Bayes loss $L_{01}^{*}$ for this problem, and compute its value by numerical integration.

[e. Optional but helpful as a sanity check] Make a plot of $pg_{+}(x)$ and $(1-p)g_{-}(x)$ on the same graph. Mark also the locations of $\mu_{\pm}, \theta_{g}, \theta_{L}, \theta_{*}^{1}$ on the graph.

[f. Optional-extra credit: Linear classification with outliers] Now “add” the outlier (100, +1) to the original data set. Recalculate $\theta_{g}$ and $\theta_{L}$ with the new data. *No derivations for this part, just numerical results OK.*

Compare with the values in b, c and explain what you observe.

$^{1}\theta_{*}$ is $\theta_{g}$ computed using the true parameters.
The logit loss

\[ L_{\text{logit}}(w) = \ln(1 + e^{-yw^Tx}), \quad x, w \in \mathbb{R}^n, \quad y = \pm 1 \]  

(6)

is the negative log-likelihood of observation \((x, y)\) under the logistic regression model \(P(y = 1|x, w) = \phi(w^Tx)\) where \(\phi\) is the logistic function.

**a.** Show that the partial derivatives \(\frac{\partial L_{\text{logit}}}{\partial w_i}, \frac{\partial L_{\text{logit}}}{\partial x_i}\) for \(L_{\text{logit}}\) in (6) can be rewritten as

\[
\frac{\partial L_{\text{logit}}}{\partial w_i} = -(1 - P(y^*|x, w))yx_i 
\]

(7)

\[
\frac{\partial L_{\text{logit}}}{\partial x_i} = -(1 - P(y^*|x, w))yw_i 
\]

(8)

where \(y^* = \frac{1+y}{2}\) (which maps \(y \in \{\pm 1\}\) to \(y^* \in \{1, 0\}\)) and \(P\) is the probabilistic model in **a.**

The elegant formulas above hold for a larger class of statistical models, called Generalized Linear Models.