STAT 535 Homework 2
Out October 19, 2022
Due October 26, 2022
©Marina Meilă
mmp@stat.washington.edu

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 $P_{XY} = P_Y P_{X|Y}$, $P_Y(+1) = 0.7$, $P_{X|Y=\pm 1} = Normal(\mu_\pm, I_2)$ 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* $\tilde{\mathcal{D}}$ 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 $\mathcal{D}_b$ of size $n = 100$ from P_{XY}

   (b) Denote by f_b the K-NN classifier based on $\mathcal{D}_b$. Calculate $\hat{y}^{ib} = f_b(\tilde{x}^i)$ for $\tilde{x}^i \in \tilde{\mathcal{D}}$ *(The predictions of f_b on test sample).*

   (c) Calculate $\hat{l}_b = \frac{1}{n}\sum_{i\in\mathcal{D}_b} \mathbf{1}_{[\hat{y}^{ib} \neq y^i]}$ for $(x^i, y^i) \in \mathcal{D}_b$. *(How well does f_b fit the training set)*

   (d) Calculate L_b the (estimated) expected loss of f_b

$$L_b \equiv L(f_b) = \frac{1}{\tilde{n}} \sum_{(\tilde{x}^i, \tilde{y}^i)\in\tilde{\mathcal{D}}} \mathbf{1}_{[f_b(\tilde{x}^i) \neq \tilde{y}^i]} \tag{1}$$

b. Calculate the average and variance of the expected losses; denote $L = $average$(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

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

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 $f(\tilde{x}^i)$.

$$V = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} p_i(1-p_i) \tag{3}$$

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

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

f. Interpret the graphs in **e.**. Which graphs informs about the variance of f, the K-NN classfier? 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 classfier? 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–NOT GRADED]

This homework will make use of the (one-dimensional) data set $\mathcal{D}$ 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

$$f_g(x) \;=\; \begin{cases} +1 & \text{if } x > \theta_g \\ -1 & \text{if } x < \theta_g \\ 0 & \text{if } x = \theta_g \end{cases}, \tag{4}$$

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

c. 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.

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

$$f_L(x) \;=\; \mathrm{sgn}(sx - \theta_L) \tag{5}$$

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

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

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

Problem 3 – Kernel regression and its bias

In this problem, the true regression function is $f(x) = x^2 + 1$, and the sampling density is $p_X \propto \frac{\alpha}{3} Normal(0, 0.3^2) + \frac{4}{3} Normal(1, 0.6^2)$, when $x \in [-1, 1]$ and 0 otherwise. The parameter α needs to be chosen so that this density integrates to 1.

The file `hw2_kr.dat` contains $n = 300$ samples from this density; denote $\mathcal{D} = \{(x^i, y^i = f(x^i), \, i = 1 : N\}$. This problem examines the empirical properties of the Nadaraya-Watson regressor with Gaussian kernel ($b(z)$ is a standard normal) and kernel width $h = 0.1$ and relates them to the known theory.

a. Give the analytic expression, then calculate the value of α.

b. Calculate the values of $\hat{y}(x)$ the kernel regressor and plot $f(x)$ and $\hat{y}(x)$ on $[-1.5, 1.5]$ on the same graph. *For the next graphs, keep the x axis of the same size* $[-1.5, 1.5]$*, so that they can be compared with this one.*

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

c. Calculate and plot the error $\hat{y} - f$.

d. Plot the data density, p_X.

e. On the next graph, plot $f', (f'')^2$ on $x \in [-1.5, 1.5]$, as well as $\frac{p'_X}{p_X}$.

f. The theoretical bias of the Nadaraya-Watson regressor is proportional (see supplementary notes on Course notes page) with bias $= f' \frac{p'_X}{p_X} + (f'')^2$. Note that this bias is the *expectation* of $\hat{y} - f$ over samples of size n.

Plot on the same graph bias and $\hat{y} - f$; rescale bias by a constant of your choice, so that the two graphs are comparable (e.g. of the same order of magnitude). Are the two graphs similar?

g. Is there a border effect at $x = +1$? Explain why or why not. Is there a border effect at $x = -1$? Explain why or why not.

h. Explain the bias observed at $x = 0$.

Problem 4 – Bayes loss

The data in Problem 2 were generated from two normal distributions with means $\mu_+ = 2, \mu_- = -1.2$, variance 1, and $p = 1/3$. Use this true data distribution and the information in Problem 2 to answer the following questions.

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

b. Then, write the expression of the Bayes classifier f^*, and Bayes loss L^*_{01} for this problem, and compute L^* by numerical integration.

Denote by $\theta_* \in \mathbb{R}$ the decision boundary of the Bayes classifier f^*. Compute the value of θ_*?

c. Make a plot of $pg_+(x)$ and $(1 - p)g_-(x)$ on the same graph. Mark also the locations of $\mu_\pm$ and θ_*, and optionally, if you have solved Problem 2, plot also θ_g, θ_L obtained from data in Problem 2.