STAT 535 Homework 2
Out October 15, 2020
Due October 22, 2020
©Marina Meilă
mmp@stat.washington.edu

Problem 1 – Basis function predictors

Let $\mathcal{B} = \{b(\,, \xi),\ \xi \in \Xi\}$ be a finite or infinite *dictionary* and define a class of predictors $\mathcal{F}$ follows $\mathcal{F} \subseteq \{f(x) = \sum_{i=1}^{M} \beta_i b(x; \xi_i),\ \xi_{1:M} \in \Xi, \beta_{1:M} \in \mathbb{R}\}$. We say that $\mathcal{F}$ is a *basis function predictor*.

Show that each of the predictor classes below can be represented as a basis function classifier by finding a suitable dictionary and explaining what the coefficients β_i should be. For each classifier below, answer if $\mathcal{F}$ contains all the possible M-terms linear combinations over the dictionary $\mathcal{B}$ or a strict subset of them.

1. Regression trees with 1 level
2. Regression trees with any fixed number of levels l
3. 2 layer neural networks with linear outputs [Optional]
4. Naive Bayes classifiers for binary classification, where $P_{X_j|y} = \mathcal{N}(\mu_{jy}, \sigma_j^2)$

Problem 2 – 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 $P_{XY} = P_Y P_{X|Y}$, $P_Y(+1) = 0.5$, $P_{X|Y=\pm 1} = Normal(\mu_\pm, I_2)$ 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 3 – Classifiers in 1 dimension

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

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

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 $\mathcal{D}$.

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 μ_+, μ_-, 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 θ_g and θ_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.

Problem 4 – Kernel regression and its bias

In this problem, the true function is $f(x) = x^2 + 1$, and the sampling density is f_X is the mixture $\frac{\alpha}{3} Normal(0, 0.3^2) + \frac{4}{3} Normal(1, 0.6^2)$, supported on $[-1, 1]$. The parameter α needs to be chosen so that this density integrates to 1; the value of α is close 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 theoretical and empirical properties of the Nadaraya-Watson regressor with Gaussian kernel ($b(z)$ is a standard normal) and kernel width $h = 0.1$.

a. Calculate the value of the parameter α.

b. Calculate the values of $\hat{y}(x)$ the kernel regressor and plot $f(x)$ on $[-1, 1]$ and $\hat{y}(x)$ on $[-1.5, 1.5]$ on the same graph.

c. Bias of $\hat{y}$. Calculate and plot on the same graph, for $x \in [-1, 1]$ the following: the bias $\hat{y} - f$ and the theoretical bias from Lecture Notes II.1, equation (8), ignoring the vanishing terms.

d. Plotting the first and second order bias terms for fixed data $\mathcal{D}$ from (7). First, plot the data density, and the x-dependent components of the bias, namely $A_1(x) = \sum_{i=1}^N w_i(x^i - x)$ and $A_2(x) = \sum_{i=1}^N w_i(x^i - x)^2$ on the same graph, $x \in [-1.5, 1.5]$.

e. On the next graph, plot f', f'' on $x \in [-1.5, 1.5]$.

f. Now finally plot the first and second order bias terms of (7), denote them $b_1, b_2(x)$ on a third plot, as well as their sum (which is the total bias estimated

[1] θ_* is θ_g computed using the true parameters.

by (7)) and the true $\hat{y} - f$. Let the x axis be $[-1.5, 1.5]$ but plot $\hat{y} - f$ only on $[-1, 1]$. *It is most illustrative if the graphs in d, e, f have the x axes aligned.*

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 $x = 0$.