COMS 4771 HW1 (Spring 2022)

Due: Feb 18, 2022 at 11:59pm

This homework is to be done **alone**. No late homeworks are allowed. To receive credit, a typesetted copy of the homework pdf must be uploaded to Gradescope by the due date. You must show your work to receive full credit. Discussing possible approaches for solutions for homework questions is encouraged on the course discussion board and with your peers, but you must write their own individual solutions and **not** share your written work/code. You must cite all resources (including online material, books, articles, help taken from/given to specific individuals, etc.) you used to complete your work.

1 Maximum Likelihood Estimation (MLE) versus Maximum a Posteriori (MAP) Estimation

Here we investigate the difference between MLE vs. MAP estimation using a specific example. Your friend gives you a coin with bias b (that is, tossing the coin turns '1' with probability b, and turns '0' with probability $1 - b$). You make n independent tosses and get the observation sequence $x_1, \ldots, x_n \in \{0, 1\}$.

(i) You want to estimate the coin's bias. What is the Maximum Likelihood Estimate (MLE) $\hat{b}$ given the observations $x_1, \ldots, x_n$?

(ii) Is your estimate from part (i) an unbiased estimator of b? How about consistent? Justify your answer.

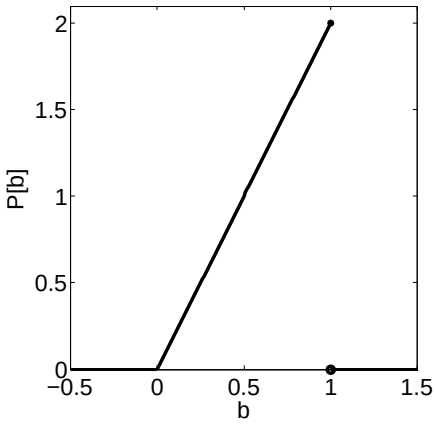
(iii) Derive a simple expression for the variance of this coin?

(iv) What is the MLE for the coin's variance?

(v) Your friend reveals to you that the coin was minted from a faulty press that biased it towards 1. Suppose the model for the faulty bias is given by the following distribution: Having this extra knowledge, what is the best estimate for the coin's bias b given the observation sequence? That is, compute: $\arg\max_b P[b \mid x_1, \ldots, x_n]$.

Note: this estimate of the coin's bias encorporates prior knowledge, and is call a MAP estimate.

(vi) When does MAP estimate equals MLE?



2 On Forecasting Product Demand

One way retail industry uses machine learning is to predict how much quantity Q of some product to they should buy to maximize their profit. The optimal quantity depends on how much demand D there is for the product as well as its cost for the retailer to buy C and its selling price P to the customer. Assuming that the demand D is distributed as $P(D)$, we can evaluate the expected profit considering two cases:

- if $D \geq Q$, then the retailer sells all Q items and make a profit $\pi = (P - C)Q$.
- but if $D < Q$, then the retailer can only sell D items at profit $(P-C)D$, but has lost $C(Q-D)$ on unsold items.

(i) What is the expected profit if the retailer buys Q items? Simplify the expression as much as possible.

(ii) By taking the derivative (wrt Q) of the above expression for expected profit, show that the optimal quantity Q^* to by satisfies $Q^* = F^{-1}(1 - (C/P))$, where F is the cdf of D. That is, the optimal Q^* is when the cumulative density (of D) equals $1 - (C/P)$.

3 Bayes Error Rate

Consider the classification problem on an arbitrary (measurable) input space X and a binary output space $Y = \{0, 1\}$. Given a joint data distribution $\mathcal{D}$ over $X \times Y$, let $g : X \to Y$ denote the Bayes classifier $g(x) := \arg\max_Y \Pr[Y|X = x]$. Define $\text{ERR}(g) := \Pr_{(x,y)\sim\mathcal{D}}[g(x) \neq y]$ as the error rate of the Bayes classifier. Prove the following statements about the error rate of g.

(i) Prove that

$$\text{ERR}(g) = \frac{1}{2} - \frac{1}{2}\,\mathbb{E}_{x\sim\mathcal{D}|_X}\Big|2\eta(x) - 1\Big|,$$

where $\mathcal{D}|_X$ denotes the marginal distribution on X, and $\eta(x) := P[Y = 1|X = x]$.

(ii) Let $p_1 := \Pr[Y = 1]$, then

$$\text{ERR}(g) = \int_{x \in X} \min \Big\{ (1-p) f_0(x), p f_1(x) \Big\} dx,$$

where f_1 and f_0 are densities of the class conditional distributions $\Pr[X|Y=1]$ and $\Pr[X|Y=0]$ respectively.

(iii) If the class priors are equal (that is, $\Pr[Y=0] = \Pr[Y=1] = 1/2$), then

$$\text{ERR}(g) = \frac{1}{2} - \frac{1}{4} \int_{x \in X} \Big| f_1(x) - f_0(x) \Big| dx,$$

where f_1 and f_0 are densities of the class conditional distributions $\Pr[X|Y=1]$ and $\Pr[X|Y=0]$ respectively.

4 A comparative study of classification performance of handwritten digits

Download the datafile `digits.mat`. This datafile contains 10,000 images (each of size 28x28 pixels = 784 dimensions) of handwritten digits along with the associated labels. Each handwritten digit belongs to one of the 10 possible categories $\{0, 1, \ldots, 9\}$. There are two variables in this datafile: (i) Variable X is a 10,000x784 data matrix, where each row is a sample image of a handwritten digit. (ii) Variable Y is the 10,000x1 label vector where the i^{th} entry indicates the label of the i^{th} sample image in X.

Special note for those who are not using Matlab: Python users can use `scipy` to read in the mat file, R users can use `R.matlab` package to read in the mat file, Julia users can use `JuliaIO/MAT.jl`. Octave users should be able to load the file directly.

To visualize this data (in Matlab): say you want to see the actual handwritten character image of the 77th datasample. You may run the following code (after the data has been loaded):

```
figure;
imagesc(1-reshape(X(77,:),[28 28])');
colormap gray;
```

To see the associated label value:

```
Y(77)
```

(i) Create a probabilistic classifier (as discussed in class) to solve the handwritten digit classification problem. The class conditional densities of your probabilistic classifier should be modeled by a Multivariate Gaussian distribution. It may help to recall that the MLE for the parameters of a Multivariate Gaussian are:

$$\vec{\mu}_{\text{ML}} := \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i$$

$$\Sigma_{\text{ML}} := \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu}_{\text{ML}})(\vec{x}_i - \vec{\mu}_{\text{ML}})^{\mathsf{T}}$$

You must submit your code to receive full credit.

(ii) Create a k-Nearest Neighbor classifier (with Euclidean distance as the metric) to solve the handwritten digit classification problem.

You must submit your code to receive full credit.

(iii) Which classifier (the one developed in Part (i) or the one developed in Part (ii)) is better? You must justify your answer with appropriate performance graphs demonstrating the superiority of one classifier over the other. Example things to consider: you should evaluate how the classifier behaves on a holdout 'test' sample for various splits of the data; how does the training sample size affects the classification performance.

(iv) As discussed in class, there are several metrics one can use in a Nearest Neighbor classification. Do a similar analysis to justify which of the three metrics: L_1, L_2 or L_∞ is better for handwritten digit classification problem.

Note: All plots, analysis and results for this question should be included in the pdf document to receive credit.