CS 536: Homework 1

Solutions

1 Bayesian Estimation (20 pts)

In the class we saw maximum likelihood estimation. Given a random variable X distributed according to $D(\theta)$ and given i.i.d. samples $X_1, X_2, \ldots X_n$ from $D(\theta)$, the MLE estimate of θ is

$$\theta_{mle} = argmax_{\hat{\theta}} P(X_1, X_2, \ldots X_n | \hat{\theta})$$

Here $P(X_1, X_2, \ldots X_n | \hat{\theta})$ is the likelihood function assuming data is generated from $\hat{\theta}$. Suppose instead we change our estimator to be

$$\theta_{new} = argmax_{\hat{\theta}} P(\hat{\theta} | X_1, X_2, \ldots X_n)$$

Using Bayes Rule this is equivalent to

$$\theta_{new} = argmax_{\hat{\theta}} P(\hat{\theta}) P(X_1, X_2, \ldots X_n | \hat{\theta})$$

In other words, we are now maximizing a weighted likelihood function, where the weight at any $\hat{\theta}$ is given by the density function $P(\hat{\theta})$. This is known as Bayesian estimation. We will later see why it's important. For now, simply view it as another way to come up with an estimator. The weight function $P(\hat{\theta})$ is commonly known as the *prior*.

1. Compute the Bayes estimate for the problem of estimating the bias of a coin. Assume that the prior $P(\hat{\theta})$ is given by the density function of a Beta distribution with parameters $\alpha, \beta > 1$[1].

   Solution: Given $X_1, X_2, \ldots X_n$, the likelihood $P(X_1, X_2, \ldots X_n | \hat{\theta}) = \prod_{i=1}^{n} (\theta)^{X_i} (1 - \theta)^{(1-X_i)}$. Also we are given that $P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$. Hence, we need to find

   $$\theta_{new} = argmax_{\theta} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \prod_{i=1}^{n} (\theta)^{X_i} (1-\theta)^{(1-X_i)}$$

   Taking log and setting derivative with respect to θ to zero gives us $\theta_{new} = \frac{\sum_i X_i + \alpha - 1}{n+\alpha+\beta-2}$

2. Compute the Bias and Variance of your estimator.

   Solution: $E[\theta_{new}] = \frac{\alpha-1+n\theta}{n+\alpha+\beta-2}$. Hence bias $= E[\theta_{new}] - \theta = \frac{\alpha-1+\theta(2-\alpha-\beta)}{n+\alpha+\beta-2}$.

   $Var(\theta_{new}) = \frac{1}{(n+\alpha+\beta-2)^2} \sum_i Var(X_i) = \frac{n\theta(1-\theta)}{(n+\alpha+\beta-2)^2}$

2 Minimax Optimality (20 pts)

Let X be a random variable distributed according to $D(\theta)$. For any estimator $\hat{\theta}(X_1, X_2, \ldots X_n)$ of θ, define the *mean squared error* as $MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$. Here the expectation is over the random samples drawn from $D(\theta)$. Define the Bias of $\hat{\theta}$ to be $Bias(\hat{\theta}) = E[(\hat{\theta} - \theta)]$.

[1] See here for how a beta distribution looks like (`https://en.wikipedia.org/wiki/Beta_distribution`)

1. Show that $MSE(\hat{\theta}, \theta) = Bias(\hat{\theta})^2 + Variance(\hat{\theta})$

   Solution: $MSE(\hat{\theta}, \theta) = E[[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - \mu + \mu - \theta)^2]$, where $\mu = E[\hat{\theta}]$. Expanding we get, $MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \mu)^2] + E[(\mu - \theta)^2] + 2E[(\hat{\theta} - \mu)(\mu - \theta)]$. First term is $E[(\hat{\theta} - \mu)^2] = Var(\hat{\theta})$. Second term is $E[(\mu - \theta)^2] = (\mu - \theta)^2 = Bias^2(\hat{\theta})$ (The expectation is over the draws of the data and μ is an expectation and hence is independent of the draws of data.). The third term is 0 since $E[(\hat{\theta} - \mu)(\mu - \theta)] = (\mu - \theta)E[(\hat{\theta} - \mu)] = 0$.

Define the maximum risk of an estimator to be $R(\hat{\theta}) = \sup_\theta MSE(\hat{\theta}, \theta)$. Given a prior $P(\cdot)$, define the Bayes risk of an estimator to be $B_P(\hat{\theta}) = \int_\theta P(\theta) MSE(\hat{\theta}, \theta)$. A minimax optimal estimator is defined to be

$$\theta^*_{minimax} = \inf_{\hat{\theta}} R(\hat{\theta})$$

Similarly, a Bayes optimal estimator with respect to a prior $P(\cdot)$ is defined to be

$$\theta^*_{Bayes} = \inf_{\hat{\theta}} B_P(\hat{\theta})$$

The above can be written as

$$\begin{aligned}
\theta^*_{Bayes} &= \inf_{\hat{\theta}} B_P(\hat{\theta}) \\
&= \inf_{\hat{\theta}} \int_\theta P(\theta) MSE(\hat{\theta}, \theta) \\
&= \inf_{\hat{\theta}} \int_\theta P(\theta) E[(\hat{\theta} - \theta)^2] \\
&= \inf_{\hat{\theta}} \int_\theta \int_{X_1, X_2, \ldots X_n} P(\theta) P(X_1, X_2, \ldots X_n | \theta) (\hat{\theta} - \theta)^2 \\
&= \inf_{\hat{\theta}} \int_\theta \int_{X_1, X_2, \ldots X_n} P(X_1, X_2, \ldots X_n) P(\theta | X_1, X_2, \ldots X_n) (\hat{\theta} - \theta)^2 \\
&= \inf_{\hat{\theta}} \int_{X_1, X_2, \ldots X_n} P(X_1, X_2, \ldots X_n) \int_\theta P(\theta | X_1, X_2, \ldots X_n) (\hat{\theta} - \theta)^2 \qquad (1)
\end{aligned}$$

2. Using the above derivation, find the Bayes optimal estimator for the problem of estimating the bias of a coin. Assume that the prior is a Beta distribution with parameter $\alpha, \beta > 1$. [Hint: Find a single estimator that minimizes the inner integral in Equation 1].

   Solution: Let's compute the Posterior distribution $P(\theta | X_1, X_2, \ldots X_n)$. By Bayes rule, this is proportional to $P(\theta) P(X_1, X_2, \ldots X_n | \theta)$. When the prior is a beta distribution we get that

   $$P(\theta | X_1, X_2, \ldots X_n) \propto \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{n - \sum_{i=1}^n X_i} \propto \theta^{\alpha + \sum_{i=1}^n X_i - 1} (1 - \theta)^{n + \beta - 1 - \sum_{i=1}^n X_i}$$

   Hence, the posterior is also a beta distribution with parameters $(\alpha + \sum_{i=1}^n X_i, n + \beta - \sum_{i=1}^n X_i)$. The inner integral in equation 1 is asking for $\hat{\theta}$ that minimizes $E[(\hat{\theta} - \theta)^2]$ where the expectation is taken over the posterior of θ. This will be minimized when $\hat{\theta} = E[\theta] = \frac{\alpha + \sum_{i=1}^n X_i}{n + \alpha + \beta}$. Since the estimator minimizes the inner integral for every $X_1, \ldots X_n$, this is a Bayes optimal estimator.

Sufficient conditions for optimality

3. Let $\hat{\theta}$ be a Bayes optimal estimator for a prior $P()$. Furthermore, suppose that

   $$MSE(\hat{\theta}, \theta) \leq B_P(\hat{\theta}), \forall \theta$$

   Show that $\hat{\theta}$ is also a minimax optimal estimator.

   Solution: Assume that $\hat{\theta}$ is not minimax optimal. Let θ' be the minimax optimal estimator. Then we have $R(\theta') < R(\hat{\theta})$. This means that

   $$MSE(\theta', \theta) < R(\hat{\theta}) \leq B_P(\hat{\theta}), \forall \theta$$

But then θ' will be the Bayes optimal estimator with respect to $P()$. Hence, we reach a contradiction.

4. Design a minimax optimal estimator for the problem of estimating the bias of a coin. [Hint: Design a Bayes estimator with a Beta prior and choose appropriate values of α and β.]

   Solution: Let's take the estimator above, $\hat{\theta} = \frac{\alpha+\sum_{i=1}^{n} X_i}{n+\alpha+\beta}$. We have $Bias(\hat{\theta}) = \frac{\alpha-\theta(\alpha+\beta)}{n+\alpha+\beta} - \theta$, and $Var(\hat{\theta}) = \frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2}$. From part 2, we want to choose α, β such that the MSE of $\hat{\theta}$ becomes independent of θ. This would imply that $MSE(\hat{\theta}, \theta) \leq R(\hat{\theta})$ and hence will make it a minimax optimal estimator as well. The values of α, β when $MSE(\hat{\theta}, \theta)$ becomes independent of θ are $\alpha = \beta = \frac{\sqrt{n}}{2}$.

3 Minimax with 0/1 error (10 pts)

Consider again the problem of estimating the bias of a coin. Suppose you are given the promise that the unknown bias p is either $1/4$ or $3/4$, i.e., $p \in \{\frac{1}{4}, \frac{3}{4}\}$. However, we define the mean squared error in different way. Given an estimator $\hat{\theta}$, define the new MSE as $MSE_{new}(\hat{\theta}, \theta) = [Pr(\hat{\theta} \neq \theta)]$. This error is often known as the $0/1$ error. Here the probability is over the random draw of i.i.d. samples. Design a minimax optimal estimator for the unknown bias under the new definition of the mean squared error. Using the machinery developed in the previous problem, show that your estimator is minimax optimal.

Solution: Let $X_1, X_2, \ldots X_m$ be the random draws of a coin. The estimator is the following: If $\sum_{i=1}^{m} X_i \geq \frac{m}{2}$, output $p = \frac{3}{4}$, else output $p = \frac{1}{4}$.

Proof of Optimality: First notice that the derivation of equation 1 and sufficient conditions of optimality from the previous problem also hold true for the new definition of the error, i.e., the $0/1$ error. Let's choose a uniform prior P, i.e., $P(1/4) = 1/2$ and $P(3/4) = 1/2$. The Bayes risk of any estimator with respect to this prior can be written as

$$
\begin{aligned}
B_P(\hat{\theta}) &= \sum_{\theta} P(\theta) MSE_{new}(\hat{\theta}, \theta) \\
&= \sum_{\theta} P(\theta) P(\hat{\theta} \neq \theta) \\
&= \frac{1}{2} P(\hat{\theta} \neq \frac{1}{4}) + \frac{1}{2} P(\hat{\theta} \neq \frac{3}{4}) \\
&= \frac{1}{2} \left[\sum_{X_1, X_2, \ldots X_m} P(X_1, X_2, \ldots X_m | \theta = \frac{1}{4}) I(\hat{\theta} \neq \theta) + \sum_{X_1, X_2, \ldots X_m} P(X_1, X_2, \ldots X_n | \theta = \frac{3}{4}) I(\hat{\theta} \neq \theta) \right] \\
&= \frac{1}{2} \left[\sum_{X_1, X_2, \ldots X_m} \left[P(X_1, X_2, \ldots X_m | \theta = \frac{1}{4}) I(\hat{\theta} \neq \theta) + P(X_1, X_2, \ldots X_m | \theta = \frac{3}{4}) I(\hat{\theta} \neq \theta) \right] \right]
\end{aligned}
\tag{2}
$$

Here, $I(\hat{\theta} \neq \theta)$ takes value 1 if $\hat{\theta} \neq \theta$, otherwise it takes value 0. For every value of $X_1, X_2, \ldots X_m$, the estimator that minimizes the innermost sum if the one that outputs $\frac{1}{4}$ if $P(X_1, X_2, \ldots X_m | \theta = \frac{1}{4}) > P(X_1, X_2, \ldots X_m | \theta = \frac{3}{4})$, and otherwise outputs $\frac{3}{4}$. We also know that $P(X_1, X_2, \ldots X_m | \theta = \frac{1}{4}) = (\frac{1}{4})^{\sum_{i=1}^{m} X_i} (\frac{3}{4})^{m - \sum_{i=1}^{m} X_i}$. Similarly, $P(X_1, X_2, \ldots X_m | \theta = \frac{3}{4}) = (\frac{3}{4})^{\sum_{i=1}^{m} X_i} (\frac{1}{4})^{m - \sum_{i=1}^{m} X_i}$. Substituting, we get that the estimator is: predict $\frac{1}{4}$ if $\sum_{i=1}^{m} X_i < \frac{m}{2}$, else predict $\frac{3}{4}$. Since this estimator minimizes the inner sum for every input , it is a Bayes optimal estimator. Its error when $\theta = \frac{1}{4}$ is $Pr(\sum_{i=1}^{m} X_i > m/2 | \theta = 1/4)$. This is the same as its error when $\theta = \frac{3}{4}$ which is $Pr(\sum_{i=1}^{m} X_i < m/2 | \theta = 3/4)$. Hence, its error is independent of p and from previous problem this makes the estimator minimax optimal as well.

4 MLE for high dimensional data (20 pts)

Fix d and let X be a $d \times 1$ vector that is distributed according to a Gaussian distribution $N(\mu, \sigma^2 I)$, where μ is the mean vector of dimension $d \times 1$, σ is a scalar and I is the $d \times d$ identity matrix. In other words, $\sigma^2 I$ is the covariance matrix of the Gaussian.

1. Find the MLE estimate μ_{MLE} of the true mean vector μ. Compute the mean squared error of your estimator.

   Solution: Let $X_1, X_2, \ldots X_n$ be the random samples from the multivariate Gaussian. Each X_i is a d-dimensional vector. $\mu_{MLE} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. $MSE(\bar{X}, \mu) = E[\|\bar{X} - \mu\|^2] = \sum_{j=1}^{d} E[(\bar{X}_j - \mu_j)^2] = \sum_{j=1}^{d}[Bias^2(\bar{X}_j) + Var(\bar{X}_j)]$. Here $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$. We have $Bias(\bar{X}_j) = 0$ and $Var(\bar{X}_j) = \frac{\sigma^2}{n}$. Hence, the MSE of the estimator is $\frac{\sigma^2 d}{n}$

2. Now suppose we want to estimate $\theta = \|\mu\|^2 = \sum_i \mu_i^2$. Compute an estimate of θ based on the MLE estimate of μ. Does the error of your estimator go to 0 as $n \to \infty$?

   Solution: $\hat{\theta} = \|\mu_{MLE}\|^2 = \|\bar{X}\|^2$. $E[\|\bar{X}\|^2] = \sum_{j=1}^{d} E[\bar{X}_j^2] = \sum_{j=1}^{d}[Var(\bar{X}_j) + \mu_j^2] = \frac{\sigma^2 d}{n} + \|\mu\|^2$. Hence $Bias(\hat{\theta}) = \frac{\sigma^2 d}{n}$. $Var(\hat{\theta}) = Var(\sum_{j=1}^{d} \bar{X}_j^2)$. Each $\bar{X}_j$ is distributed as $N(\mu_j, \frac{\sigma^2}{n})$. Hence $\sum_{j=1}^{d} \bar{X}_j^2$ is a chi-squared distribution[2] and its variance is $\frac{2d\sigma^4}{n^2} + \frac{4\sigma^2\|\mu\|^2}{n}$. The total error of the estimator is $\frac{\sigma^2 d}{n} + \frac{2d\sigma^4}{n^2} + \frac{4\sigma^2\|\mu\|^2}{n}$ and it goes to zero if d is fixed.

3. So far we assumed that d is fixed. Now suppose that d increases with n. This is known as the high dimensional setting. Specifically, assume that d is increasing with n but bounded by $d \leq 2n$. Does the error of your estimator of θ go to zero as $n, d \to \infty$? Propose a better estimator whose error indeed goes to zero with increasing n.

   Solution: The new estimator is $\hat{\theta} = \|\bar{X}\|^2 - \frac{\sigma^2 d}{n}$. Its bias is 0 and the Variance remains the same and hence its total error is $\frac{2d\sigma^4}{n^2} + \frac{4\sigma^2\|\mu\|^2}{n}$ which goes to 0 with both d and n.

   Note: If σ is unknown then one would replace the estimator with $\hat{\theta} = \|\bar{X}\|^2 - \frac{\hat{\sigma}^2 d}{n}$ where $\hat{\sigma}$ is the MLE estimate of σ.

5 Bayesian estimation for multinomial distribution (10 pts)

Let X be a discrete random variable with support in $\{0, 1, 2, \ldots k\}$. Let $p_i = P(X = i)$. We denote $\vec{p} = (p_0, p_1, \ldots p_k)$ to be the vector of probabilities.

1. Find the MLE estimate $\vec{p}_{MLE}$ for $\vec{p}$

   Solution: Let $X_1, X_2, \ldots X_m$ be random draws from the distribution of X. Let $q_j = \frac{1}{m}\sum_{i=1}^{m} I(X_i = j)$. Then $\vec{p}_{MLE} = (q_0, q_1, \ldots q_k)$.

2. Find the Bayesian estimate for $\vec{p}$ assuming that the prior $P()$ is a Dirichlet distribution with parameter vector $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots \alpha_k)$. You can assume that all the α_i's are equal.[3]

   Solution: In this case, as in Problem 1 we need to maximize the weighted likelihood function. In other words, we need to find $(\hat{p}_0, \hat{p}_1, \ldots \hat{p}_k)$ that adds up to 1 and maximizes

   $$\prod_{j=0}^{k} \hat{p}_j^{\alpha-1} \prod_{j=0}^{k} \hat{p}_j^{q_j}$$

   Set $p_0 = 1 - \sum_{j=1}^{k} \hat{p}_j$ in the above equation and set the derivates to zero. Then we get $\hat{p}_j = \frac{mq_j + \alpha - 1}{m + k\alpha}$.

[2] See here for how a chi-squared distribution looks like (`https://en.wikipedia.org/wiki/Noncentral_chi-squared_distribution`)

[3] See here for how a Dirichlet distribution looks like (`https://en.wikipedia.org/wiki/Dirichlet_distribution`)

6 Naive Bayes Classifier (20pts)

Download the Reuters dataset from the course webpage. This is a dataset of documents, where each document is talking about a certain topic. For this homework, there are only documents talking about two different topics – "earn" or "acq". Your goal is to implement a Naive Bayes classifier that can predict the topic of a given document. Each document is represented as a long vector. Each coordinate of the vector corresponds to a given word and represents the number of times the word appears in the given document. This is known as the bag of words model. There are total 5180 words in the corpus and hence each document is a vector of length 5180.

In the downloaded folder you will find:
train.csv: Training data. Each row represents a document, each column separated by commas represents features (word counts). There are 4527 documents and 5180 words.
train labels.txt: labels for the training data containing the topics for a given document.
test.csv: Test data, 1806 documents and 5180 words
test labels.txt: labels for the test data
word indices: words corresponding to the feature indices.
For your convenience we have also included a version of this dataset in .mat format, (reuters.mat) so that you can directly import it to Matlab.

1. Implement a Naive Bayes classifier to predict the topic of a given document. Use maximum likelihood estimators for computing any conditional probabilities.

   Solution: There are two ways to approach this. Both will get you credit for the problem.
   Approach 1: Let $y = 0/1$ denote the class label of a document. For each word i, estimate two probabilities p_i and q_i, where $p_i = Pr(\text{word } i \text{ appears given } y = 0)$ and $q_i = Pr(\text{word } i \text{ appears given } y = 1)$. For instance, p_i can be estimated as $\frac{\text{Total count of the word in all documents with } y=0}{\text{total number of documents with } y=0}$. Also estimate $l_0 = P(y = 0)$ and $l_1 = P(y = 1)$. Given a new document D, compute $p(y = 0|D) = l_0 \prod_i p_i$. Similarly compute $p(y = 1|D)$ and go with the larger value.
   Approach 2: For each word i estimate two vectors $\vec{p_i} = (p_{i,0}, p_{i,1}, \ldots)$ and $\vec{q_i} = (q_{i,0}, q_{i,1}, \ldots)$. Here $p_{i,k} = Pr(\text{word } i \text{ appears } k \text{ times given } y = 0)$. Given a new document D where word i appears k_i times, compute $p(y = 0|D) = l_0 \prod_i p_{i,k_i}$. Similarly compute $p(y = 1|D)$ and go with the larger value.

2. Implement another Naive Bayes classifier using Bayes estimates for conditional probabilities with a Dirichlet prior as in the previous problem. Again, you can assume that the coordinates of Dirichlet prior vector $\vec{\alpha}$ have the same value.

   Solution: If you followed approach 2 you should use a Dirichlet prior on $\vec{p_i}$ and $\vec{q_i}$ as in the previous problem. If you followed approach 1 you should use a Beta prior on p_i's and q_i's.

3. Train with different values of the prior parameter. Plot the accuracy of your classifier on the test set as a function of the prior parameter. What value of the prior gives the best accuracy? The accuracy is defined as the fraction of documents in the test set for which your method predicts the correct topic.

   Solution: With Bayesian estimates you should get very good accuracy($> 95\%$) with both the approaches.

4. Which Naive Bayes implementation(MLE vs Bayesian estimates) does better in terms of accuracy? Why?

   Solution: Bayesian estimates will do better. When using MLE estimates, in both the approaches, if a given word in the test document does not appear in the training data or does not appear a specified number of times, the corresponding $p(y|D)$ will become 0 and hence we will not be able to get an accurate estimate.

Note: We expect you to write original code. You cannot use existing libraries or implementations of the Naive Bayes classifier. Your code should explicitly contain a file named "test_run.xxx" that reads the test file, calls your classifier and outputs the topic labels (1/0) for each document (one label per line).