# CS 536: Homework 1

**Due: Feb 11, 6pm EST**

**Instructions:**
This homework is challenging and you will likely need the entire 2 weeks to finish it. You are expected to think about a problem alone for at least 24 hours before discussing it with others. Discussions are meant to help you understand the problem better. You **should not** ask someone for solutions. Be prepared to come and explain your solution to us in person if we feel the need. Submit a single .zip or tar.gz file on sakai containing your .pdf file and your code. The name of the file should be your *netid*. Late homeworks will not be accepted.

Programming: You can use any language of your choice. We prefer that you use Matlab, Python and C/C++, but it's not required. The code should be concise with necessary comments. If we can't understand what your code is trying to do, we will take points off.

## 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 $\theta$ 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].
2. Compute the Bias and Variance of your estimator.

## 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 $\theta$, 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 $MSE(\hat{\theta}, \theta) = Bias(\hat{\theta})^2 + Variance(\hat{\theta})$

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].

**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.
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 $\alpha$ and $\beta$.]

## 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.

## 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 $\mu$ is the mean vector of dimension $d \times 1$, $\sigma$ 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 $\mu_{MLE}$ of the true mean vector $\mu$. Compute the mean squared error of your estimator.
2. Now suppose we want to estimate $\theta = \|\mu\|^2 = \sum_i \mu_i^2$. Compute an estimate of $\theta$ based on the MLE estimate of $\mu$. Does the error of your estimator go to 0 as $n \to \infty$?
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 $\theta$ go to zero as $n, d \to \infty$? Propose a better estimator whose error indeed goes to zero with increasing $n$.

## 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 $\vec{p}$
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 $\alpha_i$'s are equal.[2]

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

**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).

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