Assignment Two
ECE 4200

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc).
- **The due date is 9/27/2020, 23.59.59 Eastern time**.
- Submission rules are the same as previous assignment.

Problem 1 (15 points). In class we said that for a generative model (e.g., Naive Bayes), the optimal estimator will be the maximum aposteriori probability (MAP) estimator that when given a feature $\overline{X}$, outputs the label that satisfies:

$$\arg\max_{y\in\mathcal{Y}} p(y|\overline{X}).$$

The maximum likelihood (ML) estimator outputs the label that maximizes the likelihood (probability) of the observation (which is the feature $\overline{X}$):

$$\arg\max_{y\in\mathcal{Y}} p(\overline{X}|y).$$

In this problem we will see that this is the predictor with the least error probability if the underlying data is generated from the model.

We will simplify the setting by considering a binary classification task, where the labels have two possible values, say $\mathcal{Y} = \{-1, +1\}$. Suppose the model that generates the data is $p(\overline{X}, y)$, which is **known**.

1. What is the distribution of y when we observe a feature $\overline{X}$?
2. Suppose we predict the label -1 upon seeing $\overline{X}$. Show that the probability of error is $p(y = +1|\overline{X})$.
3. Use this to argue that for any $\overline{X}$ the prediction to minimize the error probability is

   $$\max_{y\in\{-1,+1\}} p(y|\overline{X}).$$

   This shows that the MAP estimator is the optimal estimator for the binary task. This also extends to larger $\mathcal{Y}$.

4. Show that if the distribution over the labels is uniform, namely $p(y=-1)=p(y=+1)=0.5$, then the MAP estimator and ML estimator are the same.

5. Construct *any* generative model where the MAP and ML estimator are not the same.

Problem 2. (10 points). ML vs MAP and add constant smoothing. Suppose you generate n independent coin tosses using a coin with bias μ (i.e. the probability of getting a head for each toss). Let $A(n_H, n_T)$ denote the event of getting n_H heads and $n_T = n - n_H$ tails. Show the following:

1. According to maximum likelihood principle, show that your estimate for μ should be:

$$\hat{\mu} = \frac{n_H}{n_H + n_T}.$$

   Hint: Given that the bias is μ, show that:

$$p(A(n_H, n_T)|\mu) = \binom{n_H + n_T}{n_T} \mu^{n_H} (1-\mu)^{n_T}.$$

2. Consider the following generative process: First generate μ, the bias of the coin, according to the *prior* distribution $p(\mu)$. Then generate n independent coin tosses with bias μ. Show that

$$\arg\max_{\mu} p(\mu|A(n_H, n_T)) = \arg\max_{\mu} p(\mu) p(A(n_H, n_T)|\mu).$$

3. Let the prior of the bias be a *Beta* distribution, which is a distribution over $[0, 1]$

$$p(\mu) = \frac{\mu^{\alpha}(1-\mu)^{\beta}}{\int_0^1 x^{\alpha}(1-x)^{\beta} d\mu},$$

   show that:

$$\arg\max_{\mu} p(\mu|n_H, n_T) = \frac{n_H + \alpha}{n_H + \alpha + n_T + \beta}.$$

   Remark: This shows that add constant smoothing is equivalent to inducing a *Beta* distribution prior on the parameter of the generating model.

Problem 3. (10 points). Consider the Tennis data set. In this problem, you don't need the smoothing constant when estimating the prior probabilities of the labels.

1. For $\beta = 1$ (smoothing constant), write down the probabilities of all the features conditioned on the labels. The total number of probabilities you need to compute should not be more than twenty.

2. What are the prior probabilities of the labels?

3. For a new label $(Overcast, Hot, High, Strong)$, what does the Naive Bayes classifier predict for $\beta = 0, \beta = 1$, and for $\beta \to \infty$?

Problem 4. (15 points). Recall the Gaussian distribution with mean μ, and variance σ^2. The density is given by:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Given n independent samples $X_1, \ldots, X_n$ from the same Gaussian distribution with unknown mean, and variance, let μ_{ML}, and σ^2_{ML} denote the maximum likelihood estimates of mean and variance.

Hint: (1) What is the joint density $p(x_1, \ldots, x_n)$ for i.i.d Gaussian random variables $X_1, \ldots, X_n$? (2) If $f(x) > 0$, maximizing $f(x)$ is equivalent as maximizing $\log[f(x)]$.

1. Show that
$$\mu_{ML} = \frac{\sum_{i=1}^n X_i}{n}.$$
2. Show that
$$\sigma^2_{ML} = \frac{1}{n}\left(\sum_{i=1}^n (X_i - \mu_{ML})^2\right).$$
3. What is the expectation and variance of μ_{ML}?

Problem 5. (20 points). See attached Jupyter Notebook for reference.