CSE 575 Statistical Machine Learning
Midterm Exam #1
September 22, 2016

1. Personal info:
   - Name:
   - ASU ID#:
2. There should be 10 numbered pages in this exam (including this cover sheet).
3. THIS IS CLOSED BOOK EXAM!
4. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
5. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
6. You have 75 minutes.
7. Good luck!

Question	Topic	Max. score	Score
1	Beta Distribution	27	
2	Probability	20	
3	Parameter Estimation	23	
4	Linear Models	18	
5	Naïve Bayes Classifier and Logistic Regression	12	

1 [27 Points] Beta Distribution

Hint 1: For $Beta(\alpha, \beta)$ distribution, the PDF is $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ ($x \in [0, 1]$, and $B(\alpha, \beta)$ is the beta function), the mean is $\frac{\alpha}{\alpha+\beta}$, and the mode is $\frac{\alpha-1}{\alpha+\beta-2}$
Hint 2: $Uniform([0, 1])$ distribution is a special case of Beta distribution, which is $Beta(1, 1)$
Hint 3: For this question, you may need to refer to the concept of conjugate prior
Hint 4: The PDF of $Beta(4, 2)$ distribution is given in the figure below.

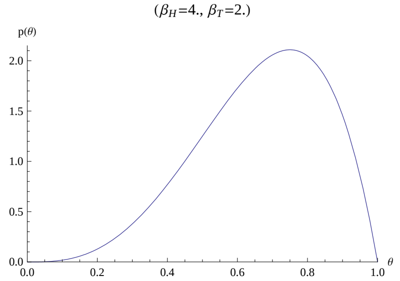

Figure 1: $Beta(4, 2)$ distribution

1. [6 points] Suppose there is a coin with unknown bias θ (probability of heads). Let n denote the total number of coin flips, n_H the total number of heads, and $n_T = n - n_H$ the total number of tails. What is the data likelihood $P(D|\theta)$? What is the log likelihood $\log P(D|\theta)$?

   Solution. The data likelihood $P(D|\theta) = \theta^{n_H}(1-\theta)^{n_T}$. The log likelihood $\log P(D|\theta) = n_H \log\theta + n_T \log(1-\theta)$.

2. [6 points] What is the MLE (maximum likelihood) estimator of θ? **Prove it.**

   Solution.

   Set
$$\frac{\partial \log P(D|\theta)}{\partial\theta} = \frac{n_H}{\theta} - \frac{n_T}{1-\theta} = 0$$

   We have the MLE estimator $\hat{\theta}$:
$$\hat{\theta} = \frac{n_H}{n_H + n_T}$$

3. [6 points] If we use $Beta(4, 2)$ as the prior for θ, and the posterior distribution $P(\theta|D)$ is $Beta(30, 40)$, in the observed data sequence, what is the total number of heads, and what is the total number of tails?

   Solution. 26 heads and 38 tails.

4. [6 points] If the posterior distribution $P(\theta|D)$ is $Beta(30, 40)$, what is the MAP (maximum a posteriori) estimator of θ? Please show the process for obtaining this estimator.

   Solution. The MAP estimator is $\frac{30-1}{30+40-2} = \frac{29}{68}$.

5. [3 points] Does there exist some value of the true parameter θ for which we would expect the MAP estimator of θ, using a $Uniform([0, 1])$ prior, to be different from the MLE estimator of θ? **Explain.**

   Solution. No. In this case, the prior's PDF is a constant $f_p(q) = 1$, so the MAP solution $\arg\max_{q\in[0,1]} f_p(q) f_x(\text{Data}; q)$ is always identical to the maximum likelihood solution $\arg\max_{q\in[0,1]} f_x(\text{Data}; q)$.

2 [20 points] Probability

1. [6 points] If A and B are **DISJOINT** events, and $P(B) > 0$, what is the value of $P(A|B)$?

   Solution. 0

2. [6 points] Suppose that the PDF of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{4}{3}(1 - x^3), & \text{for } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

   Then what is the value of $P(X < 0)$?

   Solution. 0

3. [8 points] Suppose that X is a random variable for which $E(X) = \mu$ and $Var(X) = \sigma^2$, and let c be an arbitrary constant. What is the value of $E[(X - c)^2]$?
   Hint: What is the definition of variance?

   Solution. $E[(X - c)^2] = E[X^2] - 2cE[X] + c^2 = (\mu - c)^2 + \sigma^2$

3 [23 points] Parameter Estimation

For this question, assume that $x_1, \ldots, x_N \in \mathbb{R}$ are i.i.d samples drawn from the same underlying distribution. Assume that the underlying distribution is Gaussian $N(\mu, \sigma^2)$.

1. [3 points] What is the MLE estimator of μ?

   Solution. $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

2. [4 points] Is your MLE estimator of μ a random variable? **Explain.**

   Solution. Yes. $\hat{\mu}_{MLE}$ is a function of $x_1, \ldots, x_N$. Each of them is a random variable. So $\hat{\mu}_{MLE}$ is also a random variable.

3. [6 points] Let $\hat{\mu}_{MLE}$ denote the MLE estimator of μ. Please prove that $\hat{\mu}_{MLE}$ is unbiased.
   Hint: The bias of an estimator of the parameter μ is defined to be the difference between the expected value of the estimator and μ.

   Solution. $E(\hat{\mu}_{MLE}) = E(\frac{\sum_{i=1}^{N} x_i}{N}) = \frac{1}{N} \sum_{i=1}^{N} E(x_i) = \mu$. So $\hat{\mu}_{MLE}$ is unbiased.

4. [10 points] If the true value of μ is known, then the MLE estimator of σ^2 is as follows.

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Please prove that $\hat{\sigma}^2_{MLE}$ is unbiased. Notice that this estimator is different from the one we introduced in class due to the fact that we already know the true value of μ.

Solution. $E(\hat{\sigma}^2_{MLE}) = E(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2) = \frac{1}{N}\sum_{i=1}^{N}E(x_i - \mu)^2 = \mu$. So $\hat{\sigma}^2_{MLE}$ is unbiased.

4 [18 points] Linear Models

Suppose that you have a software package for linear regression. The linear regression package takes as input a vector of responses (Y) and a matrix of features (X), where the entry $X_{i,j}$ corresponds to the i-th data point and the j-th feature for that data point and Y_i is the i-th response of the function. The linear regression package returns a vector of weights w that minimizes the sum of squared residual errors. The j-th entry of the vector, w_j is the weight applied to the j-th feature. For the following functions G_i of the input vector C_i, you should

EITHER

- specify how the response and features (Y_i and $X_{i,j}$) are calculated for the regression software package
- specify how parameters α can be obtained from the values returned by the regression software package w so that α is **the maximum likelihood estimate**

OR

- provide your reasoning for why the software can not be employed

Example. Given the function $G_i = \sum_{j=0}^{3} \alpha_j C_{i,1}^j + \epsilon_i = \alpha_0 + \alpha_1 C_{i,1} + \alpha_2 C_{i,1}^2 + \alpha_3 C_{i,1}^3 + \epsilon_i$ where $C_{i,1}$ is the first component of C_i and $\epsilon_i \sim N(0, \sigma^2)$, by setting: $X_{i,j} \leftarrow C_{i,1}^j$ for $j = 0, 1, 2, 3$ and $Y_i \leftarrow G_i$ for each i, the software package then returns $w^* = \arg\min \sum_i (y_i - \mathrm{w}_0 - \mathrm{w}_1 x_{i,1} - \mathrm{w}_2 x_{i,2} - w_3 x_{i,3})^2 = \arg\min \sum_i (G_i - \sum_{j=0}^{3} \mathrm{w}_j C_{i,1}^j)^2$. $\alpha_j \leftarrow \mathrm{w}_j$ then is the MLE for each α_j for $j = \{1, 2, 3\}$.

1. [6 points] $G_i = \alpha_1 C_{i,1}^2 e^{C_{i,2}} + \epsilon_i$ where $C_{i,2}$ is the second component of C_i and $\epsilon_i \sim N(0, \sigma^2)$.

   Solution.

   - $Y_i \leftarrow G_i$
   - $X_{i,1} \leftarrow C_{i,1}^2 e^{C_{i,2}}$
   - $\alpha_1 \leftarrow w_1$

2. [6 points] $G_i = \alpha_1 C_{i,1}^2 e^{C_{i,2}} + \epsilon_i + \gamma_i$ where $\epsilon_i \sim N(0, \sigma_1^2)$, $\gamma_i \sim N(\mu, \sigma_2^2)$, and ϵ_i and γ_i are independent. Here μ is the unknown bias and must be estimated.

   Solution. By using a few basic properties of Gaussians, we can rewrite this model as $G_i = \alpha_1 C_{i,1}^2 e^{C_{i,2}} + \mu + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_3^2)$.

   Then:

   - $Y_i \leftarrow G_i$
   - $X_{i,0} \leftarrow 1$
   - $X_{i,1} \leftarrow C_{i,1}^2 e^{C_{i,2}}$
   - $\alpha_1 \leftarrow w_1$
   - $\mu \leftarrow w_0$

3. [6 points] $G_i = \sum_j \alpha_j f_j(C_i) + \epsilon_i$ where $f_j(C_i)$ are known basis functions calculated using the input vector C_i and $\epsilon_i \sim N(0, \sigma^2)$

   Solution.

   - $Y_i \leftarrow G_i$
   - $X_{i,j} \leftarrow f_j(C_i)$
   - $\alpha_j \leftarrow w_j$

5 [12 points] Naïve Bayes Classifier and Logistic Regression

Suppose we want to train Gaussian Naïve Bayes to learn a a boolean/binary classifier: $f : X \rightarrow Y$, where X is a vector of n dimensional real-valued features: $X =< X_1, ..., X_n >$; and Y is boolean class label (i.e., $Y = 1$ or $Y = 0$). Recall that in Gaussian Naïve Bayes, we assume all X_i $(i = 1, ..., n)$ are conditionally independent given the class label Y, i.e., $P(X_i|Y = k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$ $(k = 0, 1; i = 1, ..., n)$. We also assume that $P(Y)$ follows Bernoulli$(\theta, 1 - \theta)$ (i.e., $P(Y = 1) = \theta$).

- [4 points] How many independent model parameters are there in this Gaussian Naïve Bayes classifier?

  Solution. $3n + 1$

- [8 points] Prove that the Gaussian Naïve Bayes assumption implies that $P(Y|X)$ follows the form of $P(Y = 1|X =< X_1, ..., X_n >= \frac{1}{1+exp(w_0,\sum_{i=1}^{n} w_i X_i)}$. In particular, you need to express w_i $(i = 0, ..., n)$ by the model parameters (i.e., $\theta, \mu_{ik}, \sigma_i$ $(k = 0, 1; i = 1, ..., n)$). Also, if the standard deviation σ_i also depends on the class label, i.e., $P(X_i|Y = k) \sim \mathcal{N}(\mu_{ik}, \sigma_{ik})$ $(k = 0, 1; i = 1, ..., n)$, will you still be able to convert the Gaussian Naïve Bayes classifier into the form of $P(Y = 1|X =< X_1, ..., X_n >= \frac{1}{1+exp(w_0,\sum_{i=1}^{n} w_i X_i)}$?

  Solution.

$$
\begin{aligned}
P(Y = 1|X) &= \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)} \quad \text{(Bayes Rule)} \\
&= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}} \\
&= \frac{1}{1 + exp(ln(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}))} \\
&= \frac{1}{1 + exp(ln(\frac{1-\theta}{\theta}) + \sum_i \; ln(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}))} \quad \text{(Naïve Bayes Assumption)}
\end{aligned}
$$

Since $P(X_i|Y = k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$ $(k = 0, 1; i = 1, ..., n)$, we have $P(X_i = x|Y = k) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_i^2}}$, which implies

$$
ln(\frac{P(X_i|Y = 0)}{P(X_i|Y = 1)}) = \sum_i \; (\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2})
$$

which completes the proof, with $w_0 = ln(\frac{1-\theta}{\theta}) + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$ and $w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}$.

We cannot convert GNB into the given form if the standard deviation depends on the class label.