CS570 Machine Learning PhD Qualifying Exam

Prof. Alice Oh

Spring 2013

Student ID: ______________________________, Name: ______________________________

Definitions

- Gamma function: $\Gamma(n) = (n-1)!$ if n is natural number
- Beta function: $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- Beta distribution pdf: $Beta(\theta|\alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Poisson distribution pdf: $Pois(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter.
- Gamma distribution pdf: $Gamma(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$

1. (10 points) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious desease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, and the probability of testing negative given that you don't have the disease is 0.99). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (You will get points for calculations, so please show all your steps in detail and it is not necessary to get the exact final answer.)

2. (10 points) The Poisson pmf is defined as $Pois(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter.
   (a) (5 points) Given a data set $D = \{x_1, \ldots, x_n\}$, derive the maximum-likelihood estimator (MLE) of λ (call it $\hat{\lambda}$).
   (b) (5 points) Derive the posterior $p(\lambda|D)$ with a conjugate Gamma prior.

3. (10 points) Consider a uniform distribution on the interval $[0, a]$. The density function is given by

$$p(x) = \frac{1}{a} I(x \in [0, a])$$

where $I(x)$ is indicator function, $I(x) = 1$ if x is true, else 0.

(a) (3 points) Given a data set $D = \{x_1, \ldots, x_n\}$ where $\forall x_i \geq 0$, derive the maximum-likelihood estimator (MLE) of a (call it $\hat{a}$).

(b) (3 points) What probability would the model assign to a new data point x_{n+1} using $\hat{a}$?

(c) (4 points) Do you see any problem with the above approach? Briefly suggest (in words) a better approach.

4. (10 points) Consider two-class classification problem with $N1$ points of class $C1$ and $N2$ points of class $C2$. We take D-dimensional input vector x, and project using w^T

$$y = w^T x$$

The mean vector of the two classes are given by

$$m_1 = \frac{1}{N1} \sum_{x_i \in C_1} x_i \qquad m_2 = \frac{1}{N2} \sum_{x_i \in C_2} x_i$$

We use μ to represent the projected mean. E.g. $\mu_1 = w^T m_1$ and $\mu_2 = w^T m_2$.
The within-class variance of the transformed data is given by

$$s_1^2 = \sum_{x_i \in C_1} (y_i - \mu_1)^2 \qquad s_2^2 = \sum_{x_i \in C_2} (y_i - \mu_2)^2$$

where $y_i = w^T x_i$

Show that

$$\frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

where S_B is the between-class covariance matrix given by

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$

and S_W is the total within-class covariance matrix given by

$$S_W = \sum_{x_i \in C_1} (x_i - m_1)(x_i - m_1)^T + \sum_{x_i \in C_2} (x_i - m_2)(x_i - m_2)^T$$

5. (10 points) [Decision Tree] For this question, you're going to answer a couple questions regarding the dataset shown below. You'll be trying to determine whether Andrew finds a particular type of food appealing based on the food's temperature, taste, and size.

Appealing	Temperature	Taste	Size
Yes	Hot	Salty	Small
Yes	Cold	Sweet	Large
No	Cold	Sour	Small
No	Hot	Sour	Small
Yes	Hot	Salty	Large
No	Hot	Sour	Large
No	Cold	Sweet	Small
Yes	Hot	Salty	Large

(a) (3 points) What is the initial entropy of *Appealing*?

(b) (4 points) Assume that *Taste* is chosen for the root of the decision tree. What is the information gain associated with this attribute?

(c) (4 points) Draw the full decision tree learned for this data (without any pruning).