Tutorial 2

Q1 : (5 points) :

Show that the matrix $\phi(\phi^T\phi)^{-1}\phi^T$ takes any vector v and projects it onto the space spanned by the columns of Φ. Use this result to show that the least-squares solution $w_{ML} = \phi(\phi^T\phi)^{-1}\phi^T t$ corresponds to an orthogonal projection of the vector **t** onto the manifold S (space span by column of Φ)

Q2 : (15 points) a) [7.5 points] Show that the likelihood function given by

$$p(\mathbf{X}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

When multiplied by Gamma distribution conjugate prior

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1}\exp(-b\lambda).$$ will result in a posterior of Gamma Distribution.

b) [7.5 points] Consider the Posterior distribution obtained in a) of the form $\mathrm{Gam}(\lambda|a_N,b_N)$ .Please find a_N and b_N.

Q3 : [6 Points]

(6 points) We have a training set consisting of samples and their labels. All samples come from one of two classes, 0 and 1. Samples are two dimensional vectors. The input data is the form $\{X1, X2, Y\}$ where $X1$ and $X2$ are the two values for the input vector and Y is the label for this sample.

After learning the parameters of a Naïve Bayes classifier we arrived at the following table:

Table 1: Naïve Bayes conditional probabilities

	$Y=0$	$Y=1$
$X1$	$P(X1=1\|Y=0)=1/5$	$P(X1=1\|Y=1)=3/8$
$X2$	$P(X2=1\|Y=0)=1/3$	$P(X2=1\|Y=1)=3/4$

Denote by w_1 the probability of class 1 (that is $w_1 = P(Y=1)$). If we know that the likelihood of the following two samples: {1,0,1},{0,1,0} given our Naïve Bayes model is 1/180, what is the value of w1? You do not need to derive an explicit value for w_1. It is enough to write a (correct ...) *equation* that has w_1 as the only unknown and that when solved would provide the value of w_1. Simplify as best as you can.

Q 4 : [5 Marks]. Use the following dataset
(https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/ML0101ENv3/labs/china_gdp.csv)
and perform following operations :
1) Read the dataset in python
2) Implement sigmoid function and generate the output y from the dataset x.
3) Plot the initial predictions against data points
b) [5 Marks] Try the same thing using the Gaussian Basis function.

Q 5 : [20 Points] Answer the following :

In this problem we will find the maximum likelihood estimator (MLE) and maximum a posteriori (MAP) estimator for the mean of a univariate normal distribution. Specifically, we assume we have N samples, $x_1, \ldots, x_N$ independently drawn from a normal distribution with *known* variance σ^2 and *unknown* mean μ.

1. [5 Points] Please derive the MLE estimator for the mean μ. Make sure to show all of your work.
2. [12 Points] Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2. Please show all of your work. HINT: You may want to make use of the fact that:

$$\beta^2\left(\sum_{i=1}^{N}(x_i-\mu)^2\right)+\sigma^2(\mu-\nu)^2 = \left[\mu\sqrt{N\beta^2+\sigma^2}-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^{N}x_i}{\sqrt{N\beta^2+\sigma^2}}\right]^2-\frac{\left[\sigma^2\nu+\beta^2\sum_{i=1}^{N}x_i\right]^2}{N\beta^2+\sigma^2}+\beta^2\left(\sum_{i=1}^{N}x_i^2\right)+\sigma^2\nu^2$$

3. [3 Points] Please comment on what happens to the MLE and MAP estimators as the number of samples N goes to infinity.

Q 6 : Answer the following :

Suppose that we are given an independent and identically distributed sample of n points $\{y_i\}$ where each point $y_i \sim \mathcal{N}(\mu, 1)$ is distributed according to a normal distribution with mean μ and variance 1. You are going to analyze different estimators of the mean μ.

(a) [5 points] Suppose that we use the estimator $\hat{\mu} = 1$ for the mean of the sample, ignoring the observed data when making our estimate. Give the bias and variance of this estimator $\hat{\mu}$. Explain in a sentence whether this is a good estimator in general, and give an example of when this is a good estimator.

(b) [4 points] Now suppose that we use $\hat{\mu} = y_1$ as an estimator of the mean. That is, we use the first data point in our sample to estimate the mean of the sample. Give the bias and variance of this estimator $\hat{\mu}$. Explain in a sentence or two whether this is a good estimator or not.

(c) [4 points] In the class you have seen the relationship between the MLE estimator and the least squares problem. Sometimes it is useful to use the following estimate

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n}(y_i-\mu)^2+\lambda\mu^2$$

for the mean, where the parameter $\lambda > 0$ is a known number. The estimator $\hat{\mu}$ is biased, but has lower variance than the sample mean $\bar{\mu} = n^{-1}\sum_i y_i$ which is an unbiased estimator for μ. Give the bias and variance of the estimator $\hat{\mu}$.

Q 7 :

Here are some short questions to check your basic understanding of course material.

1. [2 pts] True or False? If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions, then it will achieve zero *training error* over these training examples. Please justify your answer in one sentence.

   ★ **SOLUTION:** This statement is false since there will still be unavoidable error. If the true probability of $P(X_1 = 1, X_2 = 1|Y = 0) = 0.1$ and $P(X_1 = 1, X_2 = 1|Y = 1) = 0.2$, then we will predict $Y = 1$ if we see $X_1 = 1, X_2 = 1$. However, we will misclassify points that have $X_1 = 1, X_2 = 1, Y = 0$.

2. [2 pts] Prove that $P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$. (*Hint:* This is a two-line proof.)

   ★ **SOLUTION:** $P(X_1|X_2)P(X_2) = P(X_1 \wedge X_2) = P(X_2|X_1)P(X_1)$

3. [2 pts] True or False? After we train a logistic regression classifier, we can translate its learned weights W into the parameters of an equivalent GNB classifier for which we assume $\sigma_{ik} = \sigma_i$. Give a precise *one sentence* justification for your answer.

   ★ **SOLUTION:** This is true. Logistic regression produces a linear classification boundary and Gaussian Naive Bayes (with the equivalent variance assumption) is capable of producing any linear classification boundary. From a more mathematical perspective, we noted in the class slides (9-29-2011 lecture, page 10) that we can take a GNB classifier (with the variance assumption) and translate its parameters into the parameters of a logistic regression classifier. We saw here that setting

   $$w_i = \sum_i \left(\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right)$$

   produces a logistic regression classifier that is identical to the original GNB classifier. To go in the opposite direction, choose $\sigma_i = 1$ for convenience. Then, we need only choose μ_{i0} and μ_{i1} such that the expression on the righthand-side is w_i. This can always be done since we have two parameters we can choose.

Q 8 :

In class we discussed the fact that machine learning algorithms for function approximation are also a kind of estimator (of the unknown target function), and that errors in function approximation arise from three sources: bias, variance, and unavoidable error. In this part of the question you are going to analyze error when training Bayesian classifiers.

Suppose that Y is boolean, X is real valued, $P(Y = 1) = 1/2$ and that the class conditional distributions $P(X|Y)$ are uniform distributions with $p(X|Y = 1) = \text{uniform}[1, 4]$ and $p(X|Y = 0) = \text{uniform}[-4, -1]$. (we use uniform[a,b] to denote a uniform probability distribution between a and b, with zero probability outside the interval [a,b]).

(a) [1 point]. Plot the two class conditional probability distributions $p(X|Y = 0)$ and $p(X|Y = 1)$.

(b) [4 points]. What is the error of the optimal classifier? Note that the optimal classifier knows $P(Y = 1)$, $p(X|Y = 0)$ and $p(X|Y = 1)$ perfectly, and applies Bayes rule to classify new examples.

(c) [5 points] Suppose instead that $P(Y = 1) = 1/2$ and that the class conditional distributions are uniform distribution with $p(X|Y = 1) = \text{uniform}[0, 4]$ and $p(X|Y = 0) = \text{uniform}[-3, 1]$. What is the unavoidable error in this case? Justify your answer.

Q 9 : [10 Points] Computes the gradient of the quadratic function of x given the starting points x=[1,2,3] and then uses the result of the gradient to feed the next iterations, with new points. Prints out the result of the function at each iteration till 3rd run. Use Python script to print the results.

Q 10 : [10 Points] With given input and output relation

x = [-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1]

t = [-4.9, -3.5, -2.8, 0.8, 0.3, -1.6, -1.3, 0.5, 2.1, 2.9, 5.6]

Please fit a curve with M=4 Gaussian basis functions having unity variance.

Q 11 :

Suppose you have the following training set with three boolean input x, y and z, and a boolean output U.

x	y	z	U
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

Supppose you have to predict U using a naive Bayes classifier,

(a) *(3 points)* After learning is complete what would be the predicted probability $P(U = 0|x = 0, y = 1, z = 0)$?

$$
\begin{aligned}
& P(U = 0|x = 0, y = 1, z = 0) \\
= & \frac{P(U = 0)P(X = 0|U = 0)P(Y = 1|U = 0)P(Z = 0|U = 0)}{P(X = 0, Y = 1, Z = 0)} \\
= & \frac{P(U = 0)P(X = 0|U = 0)P(Y = 1|U = 0)P(Z = 0|U = 0)}{P(U = 0)P(X = 0, Y = 1, Z = 0|U = 0) + P(U = 1)P(X = 0, Y = 1, Z = 0|U = 1)} \\
= & \frac{8}{35} \\
= & 0.229
\end{aligned}
$$

(b) *(3 points)* Using the probabilities obtained during the Bayes Classifier training, what would be the predicted probability $P(U = 0|x = 0)$?

$P(U = 0|x = 0) = \frac{1}{2}$

In the next two parts, assume we learned a Joint Bayes Classifier. In that case...

(c) *(3 points)* What is $P(U = 0|x = 0, y = 1, z = 0)$?

$P(U = 0|x = 0, y = 1, z = 0) = 0$

(d) *(3 points)* What is $P(U = 0|x = 0)$?

$P(U = 0|x = 0) = \frac{1}{2}$

Q 12 :

(5 points) Consider a single sigmoid threshold unit with three inputs, x_1, x_2, and x_3.

$$y = g(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3) \quad \text{where} \quad g(z) = \frac{1}{1 + \exp(-z)}$$

We input values of either 0 or 1 for each of these inputs. Assign values to weights w_0, w_1, w_2 and w_3 so that the output of the sigmoid unit is greater than 0.5 if an only if (x_1 AND x_2) OR x_3.