Assignment 2

1. Point Estimation (15 pts)

The Poisson distribution is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, packet arrival density is often modeled with the Poisson distribution. If X is Poisson distributed, i.e., $X \sim Poisson(\lambda)$, its probability mass function takes the following form:

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}.$$

It can be shown that $\mathbb{E}(X) = \lambda$. Assume now we have n i.i.d. data points from $Poisson(\lambda)$: $\mathcal{D} = \{X_1, \cdots, X_n\}$. (For the purpose of this problem, you can only use the knowledge about the Poisson and Gamma distributions provided in this problem.)

(a) Show that the sample mean $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimate (MLE) of λ and it is unbiased ($\mathbb{E}\hat{\lambda} = \lambda$).

(b) Now let's be Bayesian and put a prior distribution over λ. Assuming that λ follows a Gamma distribution with the parameters (α, β), its probability density function:

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda},$$

Where $\Gamma(\alpha) = (\alpha - 1)!$ (here we assume α is a positive integer). Compute the posterior distribution over λ.

(c) Derive an analytic expression for the maximum a posterior (MAP) of λ under $Gamma(\alpha, \beta)$ prior.

2. Source of Error: Part 1 (15 pts)

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) 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 $\hat{\mu}$. Explainin a sentence whether this is a good estimator in general, and give an example of when this is a good estimator.

(b) 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 thisestimator $\hat{\mu}$. Explain in a sentence or two whether this is a good estimator or not.

(c) 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 $\hat{\mu}$.

3. Source of Error: Part 2 (15 pts)

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 $p(X|Y = 1) = uniform[1,4]$ and $p(X|Y = 0) = uniform[-4, -1]$. (we use $uniform[a, b]$ to denote a uniform probability distribution between a and b, with zero probability outside the interval $[a, b]$).

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

(b) 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. Recall that the error of a classifier is the probability that it will misclassify a new x drawn at random from $p(X)$. The error of this optimal Bayes classifier is the unavoidable error for this learning task.

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

(d) Consider again the learning task from part (a) above. Suppose we train a Gaussian Naive Bayes (GNB) classifier using n training examples for this task, where $n \to \infty$. Of course our classifier will now (incorrectly) model $p(X|Y)$ as a Gaussian distribution, so it will be biased: it cannot even represent the correct form of $p(X|Y)$ or $P(Y|X)$.
Draw again the plot you created in part (a), and add to it a sketch of the learned/estimated class conditional probability distributions the classifier will derive from the infinite training data. Write down an expression for the error of the GNB. (hint: your expression will involve integrals - please don't bother solving them).

(e) So far we have assumed infinite training data, so the only two sources of error are bias and unavoidable error. Explain in one sentences how your answer to part (d) above would change if the number of training examples was finite. Will the error increase or decrease? Which of the three possible sources of error would be present in this situation?

4. Gaussian (Naïve) Bayes and Logistic Regression (15 pts)

Recall that a generative classifier estimates $P(\boldsymbol{X}, Y) = P(Y)P(\boldsymbol{X}|Y)$, while a discriminative classifier directly estimates $P(Y|\boldsymbol{X})$. (Note that certain discriminative classifiers are non-probabilistic: they directly estimate a function $f: X \to Y$ instead of $P(Y|X)$.) For clarity, we highlight $\boldsymbol{X}$ in bold to emphasize that it usually represents a vector of multiple attributes, i.e., $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$. However, this question does not require students to derivethe answer in vector/matrix notation.

In class we have observed an interesting relationship between a discriminative classifier (logistic regression) and a generative classifier (Gaussian naive Bayes): the form of $P(Y|\boldsymbol{X})$ derived from the assumptions of a specific class of Gaussian naive Bayes classifiers is precisely the form used by logistic regression. *The derivation can be found in the required reading:* http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf.We made the following assumptions for Gaussian naive Bayes classifiers to model $P(\boldsymbol{X}, Y) = P(Y)P(\boldsymbol{X}|Y)$:

(1) Y is a boolean variable following a Bernouli distribution, with parameter $\pi = P(Y = 1)$ and thus $P(Y = 0) = 1 - \pi$.

(2) $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, where each attribute X_i is a continuous random variable. For each X_i, $P(X_i|Y = k)$ is a Gaussian distribution $\mathcal{N}(\mu_{ik}, \sigma_i)$. Note that σ_i is the standard deviation of the Gaussian distribution (and thus σ_i^2 is the variance), which does not depend on k.

(3) For all $i \neq j$, X_i and X_j are conditionally independent given Y. This is why this type of classifier is called “naive”.

We say this is a specific class of Gaussian naive Bayes classifiers because we have made an assumption that the standard deviation σ_i of $P(X_i|Y=k)$ does not depend on the value k of Y. This is not a general assumption for Gaussian naive Bayes classifiers.

Let's make our Gaussian naive Bayes classifiers a little more general by removing the assumption that the standard deviation σ_i of $P(X_i|Y=k)$ does not depend on k. As a result, for each X_i, $P(X_i|Y=k)$ is Gaussian distribution $\mathcal{N}(\mu_{ik}, \sigma_{ik})$, where $i = 1,2,\ldots,n$ and $k = 0,1$. Note that now the standard deviation σ_{ik} of $P(X_i|Y=k)$ depends on both the attribute index i and the value k of Y.

Question: is the new form of $P(Y|\boldsymbol{X})$ implied by this more general Gaussian naive Bayes classifier still the form used by logistic regression? Derive the new form of $P(Y|\boldsymbol{X})$ to prove your answer.

5. Programming (40 pts)

In this lab, please submit your code according to the following guidelines:

(a) Cross-Validation: https://qffc.uic.edu.cn/home/content/index/pid/276/cid/6530.html
Please try these three approaches holdout, K-fold and leave-p-out with the data file 2.1-Exercise.csv.
Submit 'handout.py', 'k-fold.py', and 'leave-p-out.py'

(b) Linear regression: https://qffc.uic.edu.cn/home/content/index/pid/276/cid/6541.html
Please modify linear_regression_lobf.py with the data file 2.2-Exercise.csv. For this task, take the High column values as variables and Target column for prediction.
Submit 'linear_regression_lobf.py'

(c) Naïve Bayes: https://qffc.uic.edu.cn/home/content/index/pid/276/cid/6557.html
Here the dataset 'basketball.csv' used is for basketball games and weather conditions where the target is if a basketball game is played in the given conditions or not, the dataset is very small, just containing 14 rows and 5 columns. Please modify NB.py with the data file basketball.csv, the first 4 columns are features and the last column 'play' is the target.
Submit 'NB.py'

(d) Logistic regression: https://qffc.uic.edu.cn/home/content/index/pid/276/cid/6556.html
Modify the sample code in the provided link. Use breast cancer dataset from sklearn using following code: *from sklearn.datasets import load_breast_cancer*. Use the logistic regression model in sklearn to predict the target value using all features.
Submit 'Logistic-Regression.py'