CSE 512 Machine Learning: Homework 1

Department of Computer Science
Stony Brook University

- There are 6 questions on this assignment. The last question involves coding. Do *not* attach your code to the write-up. Zip all your files, and upload the .zip file on *Blackboard*. Name your .zip file with your SBU name, e.g. vivek.zip
- The assignment is due at 5:30 PM (beginning of class) on **Thu, Feb 26, 2015**.
- Do not forget to put both your name and SBU ID on *each* page of your submission.
- If you have any questions, please direct your question first to the TA, then the instructor.
- You may *discuss* the questions with fellow students, *however* you must always write your own solutions and must acknowledge whom you discussed the question with. Do not copy from other sources, share your work with others, or search for solutions on the web.

1 Machine Learning - Problem Setup [5 points]

Given each of the tasks below, specify which kind of learning would be involved (supervised, unsupervised or reinforcement learning). Further specify what kind of a supervised task (classification, regression) or unsupervised task (clustering or density estimation) would be involved. In cases where training data might be required, give a short description of what might be the training data and labels you might use. If you believe a task can belong to more than one type, provide justification for each type.

a. Categorizing a large collection of documents to their underlying topics.

b. Identifying objects like stars, quasars etc in images of a sky survey (for example, images obtained from the Hubble Space Telescope).

c. You have estimated the incidence rates of measles for a sample set of counties in the state of New York. Based on this data you would now like to estimate the incidence rates at counties you could not explicitly sample.

2 Fundamentals [20 points]

2.1 Total Probability [5 points]

Suppose that I have two six-sided dice, one is fair and the other one is loaded – having:

$$P(x) = \begin{cases} \frac{1}{2} & x = 6 \\ \frac{1}{10} & x \in \{1, 2, 3, 4, 5\} \end{cases}$$

I will toss a coin to decide which die to roll. If the coin flip is *Heads* I will roll the fair die, otherwise the loaded one. The probability that the coin flip is *Heads* is $p \in (0, 1)$.

1. [2 pts] What is the expected value of the die roll (in terms of p)?
2. [3 pts] What is the variance of the die roll (in terms of p)?

2.2 Mixture distributions [5 points]

Here a random variable X is derived from a set of other random variables $X_i,\ i = 1 \ldots k$ each defined over the same sample space. Each of these random variables is called a component and there are k components.

We also define a distribution over the components. Let C be the random variable associated with this distribution. The value realized by C indicates which component needs to be selected.

X is realized from its components and the distribution specified by C as follows:

- First a component, say X_i, is selected according to the distribution specified by C.
- Once a component X_i is selected, its value is realized.

  In summary, $X = X_i$ when $C = i$.

Thus in the above example, if X is a random variable representing the outcome of the dice roll, it is derived from 2 components namely:

- X_1: The random variable representing the outcome of rolling the fair die.
- X_2: The random variable representing the outcome of rolling the loaded die.

C is the random variable associated with the coin toss and specifies a distribution over the 2 components.

1. [1 pts] Show the form of $P(X)$ in terms of $P(X_i)$ and $P(C)$.
2. [2 pts] Show the form of $E(X)$ in terms of $E(X|C)$. Make your answer as compact as possible.
3. [2 pts] Show the form of $\mathrm{Var}(X)$ in terms of $\mathrm{Var}(X|C)$ and $E(X|C)$. Make your answer as compact as possible.

2.3 Gaussian in High Dimensions [10 points]

A great deal of current work in machine learning is concerned with data which are in a high dimensional space (for example text documents, which may be seen as vectors in the lattice points of $\mathbf{R}^d$ where d is the number of words in the language). In this problem we will see that we must be careful when porting familiar concepts in $\mathbf{R}^3$ into higher dimensions.

Consider the $d-$dimensional Gaussian distribution:

$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$$

where $\mathbf{I}$ is the identity matrix of size $d \times d$ and $\mathbf{0}$ is the zero vector.

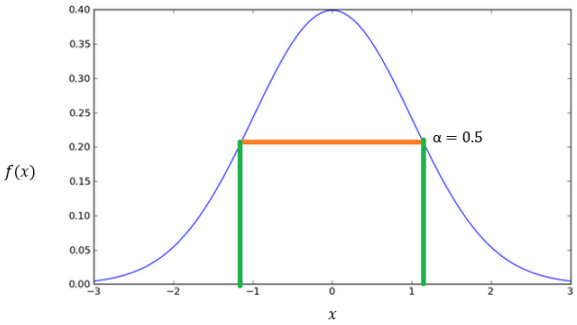
1. [2 pts] Show that the distribution of $\mathbf{x^T x}$ is the same as the distribution of $T_d = \sum_{i=1}^{d} y_i^2$ where $y_i \sim N(0, 1)$ are independent Gaussians.
2. [2 pts] Compute the mean and variance of y_i^2.
   (*Hint-1*: You may find the following useful: $E[g(z)(z - \mu)] = \sigma^2 E[g'(z)]$ where $z \sim N(\mu, \sigma^2)$.)
   (*Hint-2*: Or, you may find it handy to use the moment generating function for Gaussian distribution.)
3. [2 pts] Compute the mean and variance of T_d.
4. [2 pts] Prove that the set of points $\mathbf{x}$ with density at least α $(0 < \alpha < 1)$ times the density at the mean is given by $\mathbf{x^T x} \leq -2 \ln \alpha$.

5. [2 pts] Using the above, prove that the probability of a point x being in the region where the density is at least α times the density at the mean is given by $F_{\chi^2_d}(-2\ln\alpha)$. Here $F_{\chi^2_d}(.)$ is the cumulative distribution function of the χ^2_d distribution (called "chi-square with d degrees of freedom") Set $\alpha = 0.5$ and calculate this probability for different values of d in $[1, 10]$ (see Figure 1 to get an intuition as to what this means in 1 dimension) and plot this as a function of d. Explain the implications, in particular where do most of the points lie in high dimensions ?

   Hint: Note that the sum of squares of d independent $N(0, 1)$ draws is distributed as χ^2_d.

   Note: If you are using python you may use the function *chi2.cdf* method in the scipy.stats package to calculate the probabilities. The equivalent function in MATLAB is *chi2cdf*.

Figure 1: A standard 1-D Gaussian



3 MLE and MAP [15 points]

Brief Overview of MLE and MAP

Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) are two basic principles for learning parametric distributions. In this problem you will derive the MLE and the MAP estimates for some widely-used distributions.

Before stating the problems, we first give a brief review of MLE and MAP. Suppose we consider a family of distributions (c.d.f or p.m.f.) $F := \{f(\mathbf{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where $\mathbf{x}$ denotes the random vector, $\boldsymbol{\theta}$ denotes a vector of parameters, and Θ denotes the set of all possible values of $\boldsymbol{\theta}$. Given a set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of sample points independently drawn from some $f^* \in F$, or equivalently some $f(\mathbf{x}|\boldsymbol{\theta}^*)$ such that $\boldsymbol{\theta}^* \in \Theta$, we want to obtain an estimate of the value of $\boldsymbol{\theta}^*$. Recall that in the case of an *independently and identically distributed* (i.i.d.) sample the *log-likelihood* function is in the following form:

$$l(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \log f(\mathbf{x}_i|\boldsymbol{\theta}),$$

which is a function of $\boldsymbol{\theta}$ under some fixed sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. The MLE estimate $\hat{\boldsymbol{\theta}}_{mle}$ is then defined as follows:

1. $\hat{\boldsymbol{\theta}}_{mle} \in \Theta$.
2. $\forall \boldsymbol{\theta} \in \Theta,\ l(\boldsymbol{\theta}) \leq l(\hat{\boldsymbol{\theta}}_{mle})$.

If we have access to some prior distribution $p(\boldsymbol{\theta})$ over Θ, be it from past experiences or domain knowledge or simply belief, we can think about the *posterior* distribution over Θ:

$$q(\boldsymbol{\theta}) \;:=\; \frac{\Big(\prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta})\Big)p(\boldsymbol{\theta})}{z(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)}, \quad \text{where } z(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \;:=\; \int_{\Theta} \Big(\prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta})\Big)p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The MAP estimate $\hat{\boldsymbol{\theta}}_{map}$ is then defined as follows:

1. $\hat{\boldsymbol{\theta}}_{map} \in \Theta$.
2. $\forall \boldsymbol{\theta} \in \Theta,\ q(\boldsymbol{\theta}) \leq q(\hat{\boldsymbol{\theta}}_{map})$, or equivalently,

$$l(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \;\leq\; l(\hat{\boldsymbol{\theta}}_{map}) + \log p(\hat{\boldsymbol{\theta}}_{map}).$$

MLE and MAP for Poisson Distribution

The Poisson distribution is useful for modeling the number of events occurring within a unit time, such as the number of packets arrived at some server per minute. The probability mass function of a random variable X that follows a Poisson distribution is as follows:

$$P(X = k|\lambda) \;:=\; \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda > 0$ is the parameter of the distribution and $k \in \{0, 1, 2, \ldots\}$ indicates the number of events encountered per unit time.

1. [4 pts] Let $\{k_1, k_2, \ldots, k_n\}$ be an i.i.d. sample drawn from a Poisson distribution with parameter λ. Derive the MLE estimate $\hat{\lambda}_{mle}$ of λ based on this sample.
2. [5 pts] First, let us make an important but subtle distinction between a MLE estimate and a MLE estimator. Observe that the MLE estimate depends on the data sample. Thus different data samples will result in different MLE estimates. Thus an MLE estimate is a particular realization of a MLE estimator. The MLE estimator is thus a random variable. By convention, we use the same symbol to refer to both, with the meaning being clear from the context.

   We now explore 2 aspects of the MLE estimator:

   - First,the estimator is said to be consistent if it converges in expectation to the true parameter as the size of the training data increases.
   - Secondly the estimator should converge quickly enough to the true value as size of training data increases.

   We will show that in the case of the Poisson distribution, the MLE estimator is consistent and also asymptotically normal. Asymptotical normality implies that the MLE estimator not only converges to the true unknown parameter but converges fast enough (typically at a rate of $\frac{1}{\sqrt{n}}$).

   We will use the following facts:

   - *Law of large numbers*: If the distribution of i.i.d sample $\{X_1, X_2, \ldots, X_n\}$ is such that each X_i has finite expectation then, for any arbitrarily small $\epsilon > 0$,

   $$P(|\overline{X_n} - E(X_1)| > \epsilon) \to 0 \text{ as } n \to \infty.$$

   Here $\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ is the sample mean. Thus the sample mean converges in expectation to the true mean (population mean).

   Also note that each X_i is identically distributed and thus has the same expected value.

- *Central Limit Theorem*: If the distribution of i.i.d sample $\{X_1, X_2, \ldots, X_n\}$ is such that X_i has finite expectation $E(X_1)$ and finite variance σ^2, then

$$\sqrt{n}(\overline{X_n} - E(X_1)) \rightarrow^d N(0, \sigma^2) \text{ as } n \rightarrow \infty$$

The symbol $\rightarrow^d$ stands for "convergence in distribution". It intuitively means the random variable $\sqrt{n}(\overline{X_n} - E(X_1))$ will behave like a standard normal variable for large enough n. The sample mean $\overline{X_n}$ is said to exhibit "asymptotic normality".

Armed with these facts, show that the MLE estimator for the Poisson distribution is consistent and asymptotically normal. What is the variance of the normal distribution it converges to ?

3. [4 pt] Suppose you believe the Gamma distribution

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

is a good prior for λ, where $\Gamma(\cdot)$ is the Gamma function, and you also know the values of the two hyper-parameters[1] $\alpha > 0$ and $\beta > 0$. Derive the MAP estimate $\hat{\lambda}_{map}$.

4. [2 pt] What happens to $\hat{\lambda}_{map}$ when the sample size n goes to zero or infinity? How do they relate to the prior distribution and $\hat{\lambda}_{mle}$?

4 Naive Bayes [20 points]

1. [10 pts] Consider the learning function $\mathbf{X} \rightarrow \mathbf{Y}$, where class label $\mathbf{Y} \in \{T, F\}$, $\mathbf{X} = \langle X_1, X_2, \ldots, X_n \rangle$ where X_1 is a (1-d) boolean variable and $\{X_2, \ldots, X_n\}$ are (1-d) continues variables. Assume that for each continuous X_i, $P(X_i|Y = y)$ follows a Gaussian distribution. Besides, we also assume that the variance of X_i is independent of class.

   (a) List and give the total *number* of the parameters that you would need to estimate in order to classify a future example using a Naive Bayes classifier.

   (b) Give the formula for computing $P(Y|X)$ in terms of these parameters and feature variables X_i.

2. [10 pts] Consider a simple learning problem of determining whether Alice and Bob from CA will go to hiking or not $\mathbf{Y} : Hike \in \{T, F\}$ given the weather conditions $\mathbf{X}_1 : Sunny \in \{T, F\}$ and $\mathbf{X}_2 : Windy \in \{T, F\}$ by a Naive Bayes classifier. Using training data, we estimated the parameters $P(Hike) = 0.5$, $P(Sunny|Hike) = 0.8$, $P(Sunny|\neg Hike) = 0.65$, $P(Windy|Hike) = 0.4$ and $P(Windy|\neg Hike) = 0.5$. Assume that the *true* distribution of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{Y}$ satisfies the Naive Bayes assumption of conditional independence with the above parameters.

   (a) Assume $Sunny$ and $Windy$ are truly independent given $Hike$. Write down the Naive Bayes decision rule for this problem using *both* attributes $Sunny$ and $Windy$.

   (b) Given the decision rule above, what is the expected *error rate* of the Naive Bayes classifier? (The expected error rate is the probability that each class generates an observation where the decision rule is incorrect.)

   (c) What is the joint probability that Alice and Bob go to hiking and the weather is sunny and windy, that is $P(Sunny, Windy, Hike)$?

[1] It is common to refer to parameters in a prior distribution as hyper-parameters.

Next, suppose that we gather more information about weather conditions and introduce a new feature denoting whether the weather is $\mathbf{X}_3$: *Rainy* or not. **Assume** that each day the weather in CA can be **either** *Rainy* **or** *Sunny*. That is, it can not be both *Sunny* **and** *Rainy* (similarly, it can not be $\neg$*Sunny* **and** $\neg$*Rainy*).

(d) In the above new case, are any of the Naive Bayes assumptions violated? Why (not)? What is the joint probability that Alice and Bob go to hiking and the weather is sunny, windy and not rainy, that is $P(Sunny, Windy, \neg Rainy, Hike)$?

(e) What is the expected error rate when the Naive Bayes classifier uses all *three* attributes? Does the performance of Naive Bayes improve by observing the new attribute *Rainy*? Explain why.

5 Decision Trees [15 points]

5.1 ID3 and KL Divergence [9 points]

Consider the following set of training examples for the unknown target function $< X_1, X_2 > \rightarrow Y$. Each row indicates the values observed, and how many times that set of values was observed. For example, $(+, T, T)$ was observed 3 times, while $(-, T, T)$ was never observed.

Y	X_1	X_2	Count
+	T	T	3
+	T	F	5
+	F	T	5
+	F	F	2
-	T	T	0
-	T	F	2
-	F	T	3
-	F	F	5

Table 1:

1. [2 pts] Compute the sample entropy $H(Y)$ for this training data (with logarithms base 2)?
2. [4 pts] What are the information gains $IG(X_1) \equiv H(Y) - H(Y|X_1)$ and $IG(X_2) \equiv H(Y) - H(Y|X_2)$ for this sample of training data?
3. [3 pts] Draw the decision tree that would be learned by ID3 (without postpruning) from this sample of training data.

5.2 Information Gain and Entropy [6 points]

When we discussed learning decision trees in class, we chose the next attribute to split on by choosing the one with maximum information gain, which was defined in terms of entropy. To further our understanding of information gain, we will explore its connection to *KL-divergence*, an important concept in information theory and machine learning. For more on these concepts, refer to Section 1.6 in Bishop.

The KL-divergence from a distribution $p(x)$ to a distribution $q(x)$ can be thought of as a measure of dissimilarity from P to Q:

$$KL(p||q) = -\sum p(x)\ log_2 \frac{q(x)}{p(x)}$$

We can define information gain as the KL-divergence from the observed joint distribution of X and Y to the product of their observed marginals.

$$IG(x,y) \equiv KL\left(p(x,y)||p(x)p(y)\right) = -\sum_x \sum_y p(x,y)\ log_2\left(\frac{p(x)p(y)}{p(x,y)}\right)$$

When the information gain is high, it indicates that adding a split to the decision tree will give a more accurate model.

1. [3 pts] Show that definition of information gain above is equivalent to the one given in class. That is, show that $IG(x,y) = H[x] - H[x|y] = H[y] - H[y|x]$, starting from the definition in terms of KL-divergence.
2. [3 pts] In light of this observation, how can we interpret information gain in terms of dependencies between random variables? A brief answer will suffice.

6 Naive Bayes vs Logistic Regression [25 points]

In this problem you will implement Naive Bayes and Logistic Regression, then compare their performance on a document classification task. The data for this task is taken from the 20 Newsgroups data set[2], and is available at `http://www.cs.stonybrook.edu/~leman/courses/15CSE512/hws/hw1-data.tar.gz`. The included `README.txt` describes the data set and file format.

Our Naive Bayes model will use the bag-of-words assumption. This model assumes that each word in a document is drawn independently from a multinomial distribution over possible words. (A multinomial distribution is a generalization of a Bernoulli distribution to multiple values.) Although this model ignores the ordering of words in a document, it works surprisingly well for a number of tasks. We number the words in our vocabulary from 1 to m, where m is the total number of distinct words in all of the documents. Documents from class y are drawn from a class-specific multinomial distribution parameterized by θ_y. θ_y is a vector, where $\theta_{y,i}$ is the probability of drawing word i and $\sum_{i=1}^{m} \theta_{y,i} = 1$. Therefore, the class-conditional probability of drawing document x from our Naive Bayes model is $P(X = x|Y = y) = \prod_{i=1}^{m} (\theta_{y,i})^{\text{count}_i(x)}$, where $\text{count}_i(x)$ is the number of times word i appears in x.

1. [6 pts] Provide high-level descriptions of the Naive Bayes and Logistic Regression algorithms. Be sure to describe how to estimate the model parameters and how to classify a new example.
2. [4 pts] Imagine that a certain word is never observed in the training data, but occurs in a test instance. What will happen when our Naive Bayes classifier predicts the probability of the this test instance? Explain why this situation is undesirable. Will logistic regression have a similar problem? Why or why not?

   Add-one smoothing is one way to avoid this problem with our Naive Bayes classifier. This technique pretends that every word occurs one additional time in the training data, which eliminates zero counts in the estimated parameters of the model. For a set of documents $C = x^1, ..., x^n$, the add-one smoothing parameter estimate is $\hat{\theta}_i = \frac{1+\sum_{j=1}^{n} \text{count}_i(x^j)}{D+m}$, where D is the total number of words in C (i.e., $D = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{count}_i(x^j)$). Empirically, add-one smoothing often improves classification performance when data counts are sparse.
3. [12 pts] Implement Logistic Regression and Naive Bayes. Use add-one smoothing when estimating the parameters of your Naive Bayes classifier. For logistic regression, we found that a step size around .0001 worked well. Train both models on the provided training data and predict the labels of the test

[2]Full version available from http://people.csail.mit.edu/jrennie/20Newsgroups/

data. Report the training and test error of both models. Submit your code electronically on *Blackboard* under SafeAssignments. You do *not* need to include a hard copy of your code along with your HW submission.

4. [3 pts] Which model performs better on this task? Why do you think this is the case?