10-701/15-781, Machine Learning: Homework 1

Aarti Singh
Carnegie Mellon University

- The assignment is due at 10:30 am (beginning of class) on **Mon, Sept 27, 2010**.
- Separate you answers into five parts, one for each TA, and put them into 5 piles at the table in front of the class. Don't forget to put both your name and a TA's name on each part.
- If you have question about any part, please direct your question to the respective TA who design the part.

1 Machine Learning - Problem Setup [Min Chi, 15 points]

In online debate forums, people debate issues, express their preferences, and argue why their viewpoint is right. For example, a debate can be "which mobile phone is better: iPhone or Blackberry," or "which OS is better: Windows vs. Linux vs. Mac?" Given a debate forum, how can you use machine learning to:

a. Detect the hot debate topics.

b. Identify the points of contention within the debate.

c. For a given topic, recognize which stance a person is taking in an online debate posting.

For each of the task above, please specify what type of machine learning problem it is (regression, classification, density estimation, etc). Identify what will be the training data, features and labels (if any), and what would be the output of the algorithm.

2 Probability [Rob Hall, 20 points]

2.1 Conditional Probability and the Chain Rule [5 points]

Recall the definition of a conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

1. Prove that $P(A \cap B \cap C) = P(A|B,C)P(B|C)P(C)$
2. Suppose we play a game where I present to you three doors, one of which has a prize behind. The doors are closed and so you choose a door which you think conceals the prize. After you make your choice, I open one of the two doors you didn't pick, and reveal that the prize wasn't there (note that I can always do this). Then I give you the choice whether to stick with your current door, or switch to the remaining un-opened door. What should you do to have the highest probability of winning the prize? (Hint: consider the event that your initial door conceals the prize, then consider the probability that you win given that you decide to switch or not).

2.2 Total Probability [5points]

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. What is the expectation of the die roll (in terms of p).
2. What is the variance of the die roll (in terms of p).

   Something commonly used in statistics and machine learning are so called "mixture models" which may be seen as a generalization of the above scenario. For some sample space we have several distributions $P_i(X) = P(X|C = i)$, $i = 1 \dots k$ (e.g., the two dice from above). We also have a distribution over these "components" $P(C = i)$ (e.g., the coin toss, where C is a binary RV).

3. Show the form of $P(X)$ in terms of $P_i(X)$ and $P(C)$.
4. Show the form of $E(X)$ in terms of $E(X|C)$ make your answer as compact as possible.
5. 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 from the loving embrace of $\mathbf{R}^3$ into higher dimensions.

Consider the $d-$dimensional Gaussian distribution:

$$x \sim N(0^{d\times 1}, I^{d\times d})$$

where $I^{d\times d}$ is the identity matrix of size $d \times d$.

1. Show that the distribution of $\|x\|_2^2$ 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.
2. Compute the mean and variance of y_i^2. (Hint: You may find the following useful:

   $$E[g(z)(z - \mu)] = \sigma^2 E[g'(z)]$$

   where $z \sim N(\mu, \sigma^2)$. Here g denotes a function and g' denotes the derivative of g. Note that this is called "Stein's Lemma.")
3. Compute the mean and variance of T_d.
4. Show that $P(d - 10\sqrt{2d} \leq T_d \leq d + 10\sqrt{2d}) \geq 0.99$, for suitably large d. To assist with this problem you may use Chebyshev's inequality, which is:

   $$P(|X - EX| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

5. Prove that your answer above implies that $P(\sqrt{d} - 10 \leq \sqrt{T_d} \leq \sqrt{d} + 10) \geq 0.99$.

See that although the interval in the last expression is moving out towards infinity, its length is bounded. We may see then that in high dimensions, the norm of a gaussian vector is highly likely to be contained in this small interval. Therefore we may see that in high dimension, the gaussian distribution is more like a "shell" than a sphere.

3 MLE and MAP [TK, 20 points + 5 points (bonus)]

Maximum Likelihood Estimation (MLE) and Maximum A Posterior (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,

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

3.1 [10 points] 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 Poisson distribution is as follows:

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

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

1. [2 points] Let $\{k_1, k_2, \dots, k_n\}$ 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. [4 points] Let K be a random variable following a Poisson distribution with parameter λ. What is its mean $E[K]$ and variance $\text{Var}[K]$. Since $\hat{\lambda}_{mle}$ depends on the sample used for estimation, it is also a random variable. Derive the mean and the variance of $\hat{\lambda}_{mle}$, and compare them with $E[K]$ and $\text{Var}[K]$. What do you find?
3. [2 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] $\alpha > 1$ 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 $\hat{\lambda}_{mle}$?

3.2 [10 points] Multivariate Gaussian Distribution

The density function of a p-dimensional Gaussian distribution is as follows:

$$N(\mathbf{x}|\boldsymbol{\mu}, \Lambda^{-1}) \;:=\; \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\Lambda(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{p/2}\sqrt{|\Lambda^{-1}|}},$$

where Λ is the inverse of the covariance matrix, or the so-called *precision* matrix. Let $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be an i.i.d. sample from a p-dimensional Gaussian distribution.

1. [4 points] Suppose that $n \gg p$. Derive the MLE estimates $\hat{\boldsymbol{\mu}}_{mle}$ and $\hat{\Lambda}_{mle}$.
2. [4 points] Suppose you believe the following distribution[2]

$$gw(\boldsymbol{\mu}, \Lambda) \;:=\; N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (s\Lambda)^{-1})W(\Lambda|V, \nu)$$

is a good prior for $\boldsymbol{\mu}$ and Λ, where

$$W(\Lambda|V, \nu) \;:=\; \frac{|\Lambda|^{(\nu-p-1)/2}}{Z(V, \nu)} \exp\left(-\frac{\text{tr}(V^{-1}\Lambda)}{2}\right)$$

with $\text{tr}(\cdot)$ being the trace of a square matrix and $Z(V, \nu)$ the normalization term. You also know the values of the hyper-parameters $\boldsymbol{\mu}_0 \in \mathcal{R}^p, s > 0, \nu > p+1$, and $V \in \mathcal{R}^{p \times p}$ being positive definite. Derive the MAP estimates $\hat{\boldsymbol{\mu}}_{map}$ and $\hat{\Lambda}_{map}$.
3. [2 points] Again, what happens to $\hat{\boldsymbol{\mu}}_{map}$ and $\hat{\Lambda}_{map}$ when n goes to zero or infinity? How do they relate to the prior distribution and the MLE estimates?

3.3 [5 points Bonus] Existence and Uniqueness of MLE

It is known that MLEs do not always exist. Even if they do, they may not be unique.

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

[2] It is sometimes referred to as the Gaussian-Wishart prior.

1. [2 points] Give an example where MLEs do not exist. Please specify the family of distributions being considered, and the kind of samples on which MLEs are not well-defined.
2. [2 points] Give an example where MLEs exist but are not unique. Please specify the family of distributions being considered, and the kind of samples from which multiple MLEs can be found.
3. [1 pt] By finding the two examples as described above, hopefully you have gained some intuition on the properties of the log-likelihood that are crucial to the existence and uniqueness of MLE. What are those properties?

4 Naive Bayes [Leman, 20 points]

1. **[5 points]** Consider the learning function $\mathbf{X} \to \mathbf{Y}$, where class label $\mathbf{Y} \in \{T, F\}$, $\mathbf{X} = \langle X_1, X_2, \ldots, X_d \rangle$ where X_1 is a boolean variable and $\{X_2, \ldots, X_d\}$ are continuous variables. Assume that for each continuous X_i, $P(X_i|Y = y)$ follows a Gaussian distribution. 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. Give the formula for computing $P(Y|X)$ in terms of these parameters and feature variables X_i.
2. **[15 points]** 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 $P(Hike) = 0.5$, $P(Sunny|Hike) = 0.8$, $P(Sunny|\neg Hike) = 0.7$, $P(Windy|Hike) = 0.4$ 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 $P(Sunny, Windy, Hike)$?

   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 Naive Bayes vs Logistic Regression [Jayant, 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[3], and is available from (http://www.cs.cmu.edu/∼aarti/Class/10701/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 $P(X = x|Y = y) = \prod_{i=1}^{m} (\theta_{y,i})^{\mathrm{count}_i(x)}$, where $\mathrm{count}_i(x)$ is the number of times word i appears in x.

1. **[6 points]** 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 points]** 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 $\hat{\theta}_i = \frac{1+\sum_{j=1}^{n} \mathrm{count}_i(x^j)}{D+m}$, where D is the total number of words in C (i.e., $D = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{count}_i(x^j)$). Empirically, add-one smoothing often improves classification performance when data counts are sparse.

3. **[12 points]** 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 data. Report the training and test error of both models. Submit your code along with your homework.

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

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