CS 5350/6350: Machine Learining Spring 2015

Homework 6

Handed out: Apr 13, 2015
Due date: Apr 27, 2015

General Instructions

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down (and program) your own solution. Please keep the class collaboration policy in mind.
- Feel free ask questions about the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. Your assignment should be **no more than 10 pages**. X points will be deducted if your submission is X pages beyond the page limit.
- Handwritten solutions will not be accepted.
- The homework is due by midnight of the due date. Please submit the homework as a pdf on Canvas.

1 Naive Bayes Classification

In the class, we saw how we can build a naive Bayes classifier for discrete variables. In this question, you will explore the case when features are not discrete. Suppose instances, represented by $\mathbf{x}$, are d dimensional real vectors and the labels, represented by y, are either 0 or 1.

Recall from class that the naive Bayes classifier assumes that all features are conditionally independent of each other, given the class label. That is

$$p(\mathbf{x}|y) = \prod_j p(x_j|y)$$

Now, each x_j is a real valued feature. Suppose we assume that these drawn from a class specific normal distribution. That is,

1. When $y = 0$, each x_j is drawn from a normal distribution with mean $\mu_{0,j}$ and standard deviation $\sigma_{0,j}$, and
2. When $y = 1$, each x_j is drawn from a normal distribution with mean $\mu_{1,j}$ and standard deviation $\sigma_{1,j}$

Now, suppose we have a training set $S = \{(\mathbf{x}_i, y_i)\}$ with m examples and we wish to learn the parameters of the classifier, namely the prior $p = P(y = 1)$ and the μ's and the σ's. For brevity, let the symbol θ denote all these parameters together.

(a) Write down $P(S|\theta)$, the likelihood of the data in terms of the parameters. Write down the log-likelihood.

(b) What is the prior probability p? You can derive this by taking the derivative of the log-likelihood with respect to the prior and setting it to zero.

(c) By taking the derivative of the log-likelihood with respect to the μ_j's and the σ_j's, derive expressions for the μ_j's and the σ_j's.

2 Logistic Regression

We looked maximum a posteriori learning of the logistic regression classifier in class. In particular, we showed that learning the classifier is equivalent to the following optimization problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) + \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

In this question, you will derive the stochastic gradient descent algorithm for the logistic regression classifier.

(a) What is the derivative of the function $\log\left(1 + \exp(-y \mathbf{w}_i^T \mathbf{x}_i)\right)$ with respect to the weight vector?

(b) The inner most step in the SGD algorithm is the gradient update where we use a single example instead of the entire dataset to compute the gradient. Write down the objective where the entire dataset is composed of a single example, say $(\mathbf{x}_i, y_i)$. Derive the gradient of this objective with respect to the weight vector.

(c) Write down the pseudo code for the stochastic gradient algorithm using the gradient from part (b) above.

3 The EM algorithm

The two local newspapers The Times and The Gazette publish n articles everyday. The article length in the newspapers is distributed based on the Exponential Distribution with parameter λ. That is, for an non-negative integer x:

$$P(wordcount = x|\lambda) = \lambda e^{-\lambda x}.$$

with parameters λ_T, λ_G for the Times and the Gazette, respectively.

(**Note**: Technically, using the exponential distribution is not correct here because the exponential distribution applies to real valued random variables, whereas here, the word counts can only be integers. However, for simplicity, we will use the exponential distribution instead of, say a Poisson.)

(a) Given an issue of one of the newspapers $(x_1, \ldots x_n)$, where x_i denotes the length of the ith article, what is the most likely value of λ?

(b) Assume now that you are given a collection of m issues $\{(x_1, \ldots x_n)\}_1^m$ but you do not know which issue is a Times and which is a Gazette issue. Assume that the probability of a issue is generated from the Times is η. In other words, it means that the probability that a issue is generated from the Gazette is $1 - \eta$.

Explain the generative model that governs the generation of this data collection. In doing so, name the parameters that are required in order to fully specify the model.

(c) Assume that you are given the parameters of the model described above. How would you use it to cluster issues to two groups, the Times issues and the Gazette issues?

(d) Given the collection of m issues without labels of which newspaper they came from, derive the update rule of the EM algorithm. Show all of your work.