CSCI 4705 Artificial Intelligence
Homework 1
Deadline: March 11

PART I Programming Assignment

Introduction

In the past two weeks, we have covered probabilistic models for classification tasks. This assignment will focus on implementing one of probabilistic model – Naïve Bayes Classification. In order to assess the effectiveness of this method, you will be training and testing your code on real datasets. We will provide some methods that can read training and testing datasets and store the information for you (see download link on website) (shown in the zip file).

Datasets

The provided datasets all come from the UCI machine-learning repository under classification datasets (http://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table). I have selected three datasets that I believe will give a diversity of results.

Congressional Voting Records Dataset

This dataset (http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records) classifies congressmen into democrats and republicans based on their voting record. This is a somewhat easy problem to solve, and you should achieve fairly high accuracy with even simple method of classification (e.g. Naïve Bayes Classification). This is a good dataset to use for debugging purposes.

MONKS Problems Dataset

This dataset (http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems) has arbitrary attributes and uses 0 and 1 as labels. This dataset was generated in order to be a difficult problem to solve, and was used for a learning algorithm competition. There are three problems included in this dataset, please be sure to test your algorithm on all three! I don't expect great performance on this dataset because by its nature it is a difficult problem to solve. Note that some attributes have different ranges of values than others.

Iris Dataset

This dataset (http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems) classifies flowers into different classes of iris according to the size of different features. This is a

classic classification problem because it is a mix of linearly separable and inseparable classes.

Getting Started

A good first step would be to examine the provided code scaffolding and start working with numpy (http://www.numpy.org). Create a node class and examine what data you would need to store for the method. If you don't have much experience programming data structures in python or just need help understanding how these methods work, please look at the file (classifier.py) in the zip file and check the resources through the website.

Submission Requirements

Send me an email (yzhu@hpu.edu) containing the following:

- README that includes a brief description of each class and the breakdown of work between partners. If there are any known bugs at the time of submission, please be sure to include these and a breakdown of debugging steps taken to ensure we can give you as much partial credit as possible (major bugs that aren't explained will make you lose credit for that particular method).
- All code needed to call the functions provided by our scaffolding to train and test your methods on the provided datasets. You are allowed to modify the code I provide (as long as the method headers are unchanged), but in that case please be sure to include it!
- A pdf or other text file (pdf preferred) that reports your accuracy for each dataset.

PART II Written Assignment

P1. Expectation and Variance

Recall that the expectation and variance of a continuous random variable X are given by

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$
$$Var(X) = E[(X - E[X])^2]$$

where $p(x)$ is the probability density function. Let X and Y be independent continuous random variables and let $a \in \mathbb{R}$. Using the above definitions, derive the expectation and variance of the following random variables in terms of $E[X]$, $E[Y]$, $Var[X]$, and $Var[Y]$:

a) $X + a$
b) aX
c) $X + Y$

P2. Naïve Bayes Classifiers (Theory)

In naïve Bayes, we assume that the presence of a particular feature of a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be a watermelon if it is green, round, and more than 10 pounds. We will consider all these three features as independent to each other in naïve Bayes.

Let our features $x_i, i \in [1, d]$ be binary valued and have d dimensions, i.e. $x_i \in \{0,1\}$ and our input feature vector $x = [x_1, x_2, \cdots, x_d]^T$. For each training sample, our target value $y \in \{0, 1\}$ is also a binary-valued variable. Then our model is parameterized by $\phi_{i|y=0} = p(x_i = 1|y = 0)$, $\phi_{i|y=1} = p(x_i = 1|y = 1)$, and $\phi_y = p(y = 1)$, and

$$p(y) = (\phi_y)^y (1 - \phi_y)^{1-y}$$

$$p(x|y = 0) = \prod_{i=1}^{d} p(x_i|y = 0) = \prod_{i=1}^{d} (\phi_{i|y=0})^{x_i} (1 - \phi_{i|y=0})^{1-x_i}$$

$$p(x|y = 1) = \prod_{i=1}^{d} p(x_i|y = 1) = \prod_{i=1}^{d} (\phi_{i|y=1})^{x_i} (1 - \phi_{i|y=1})^{1-x_i}$$

a) Write down the joint log-likelihood function $l(\theta) = \log \prod_{n=1}^{N} p(x^{(n)}, y^{(n)}; \theta)$ in terms of the model parameters given above. $x^{(n)}$ means the nth data point, and θ represents all the parameters, i.e. $\{\phi_y, \phi_{i|y=0}, \phi_{i|y=1}, i = 1,2, \cdots d\}$.

b) Estimate the parameters using maximum likelihood, i.e. find the solution for parameters $\phi_y, \phi_{i|y=0}$, and $\phi_{i|y=1}$.

c) When a new sample point x comes, we make the prediction based on the most likely class estimate generated by our model. Show that the hypothesis returned by Naive Bayes is linear, i.e. if $p(y = 0|x)$ and $p(y = 1|x)$ are the class probabilities returned by our model, show that there exist some α so that

$$p(y = 1|x) \geq p(y = 0|x) \text{ if and only if } \alpha^T \tilde{x} \geq 0$$

where $\alpha = [\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_d]^T$ and $\tilde{x} = [1, x_1, x_2, \cdots, x_d]^T$.

P3. Naive Bayes with discrete features

Consider the following data set on lung diseases. Your goal is to build a Naïve Bayes classifier that predicts whether a person have Bronchitis or Tuberculosis, given his/her symptoms.

Disease	X-ray Shadow	Dyspnea	Lung Inflammation
Bronchitis	Yes	Yes	Yes
Bronchitis	Yes	Yes	Yes
Bronchitis	No	No	Yes
Tuberculosis	No	No	Yes
Tuberculosis	Yes	Yes	No
Tuberculosis	Yes	No	No

1. List the distributions that would be learned if you use MLE to estimate a Naïve Bayes model from this data (e.g. *P(Dyspnea|Bronchitis)*). Include all of the estimated probabilities for each distribution. Show your work.

2. Base on your learned Naïve Bayes model, diagnose a patient with the following symptoms. Show your work.

X-ray Shadow	Dyspnea	Lung Inflammation
Yes	No	Yes