Problem assignment 5
Due: Thursday, February 20, 2020

In this assignment we continue our investigation of the "Pima" dataset. As in the previous assignment, you can download the dataset (pima.txt) and its description (pima_desc.txt) from the course web page. In addition to the complete dataset pima.txt, you have pima_train.txt and pima_test.txt you will need to use for training and testing purposes. The dataset has been obtained from the UC Irvine machine learning repository: http://www1.ics.uci.edu/~mlearn/MLRepository.html.

Problem 1. Naive Bayes model

The Naive Bayes model defines a generative classifier model in which all features are independent given the class label. In such a case the class-conditional densities over many input variables can be decomposed into a set of independent class-conditional densities, one for every input variable. For example, the conditional probability of an input \( x = \{x_1, x_2, \cdots, x_d\} \) given class 1 in the Naive Bayes model is decomposed as:

\[
p(x|y = 1) = \prod_{i=1}^{d} p(x_i|y = 1).
\]

One important concern is the choice of an appropriate parameterization of class-conditional densities. Typically we do not choose the distributions arbitrarily, instead we want to make a good educated guess. Exploratory data analysis can help us greatly to recognize types of densities that appear to match the data the best.

Problem 2.1. Exploratory data analysis

We have performed the exploratory analysis of the Pima dataset in Problem set 1. Here we reuse the programs created there and apply them to study the density models we choose to parameterize our Naive Bayes model.

Part a. Write and submit a program (main1_1.m) that:
Divides "pima.txt" data into two subsets - one with all examples with class "0", and another with all examples with class "1".

Analyzes examples in two subsets using histograms. Histograms should give you more information about the shape of the distribution of attributes. You can use the function histogram_analysis.m for this purpose.

Part b. What distribution/density would you use to fit the values of attributes 1 to 8 in the pima dataset? Choices one typically considers are Bernoulli, Binomial, Multinomial, Normal, Poisson, Gamma, exponential distributions.

Problem 2.2. Learning of the Naive Bayes classifier

The learning of the Naive Bayes model corresponds to the estimation of parameters of class-conditional distributions \( p(x_i|y = 1), p(x_i|y = 0) \) for all input components \( i \) from data and estimation of class priors \( p(y = 1), p(y = 0) \). Thus, the learning boils down to a number of 'smaller' density estimation problems.

Assume that class-conditional densities for pima dataset have the following form:

- Class-conditionals for inputs \([1 5 7 8]\) take the form of exponential distribution. The exponential distribution is defined as:
  \[
p(x|\mu) = \frac{1}{\mu} \exp \left( -\frac{x}{\mu} \right),
\]
  where \( \mu \) is the parameter. (Exponential distribution is a special case of the Gamma distribution and belongs to the exponential family).

- Class-conditionals for inputs \([2 3 4 6]\) follow univariate normal distributions:
  \[
p(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right),
\]
  with mean and standard deviation being the two parameters.

In addition assume that priors on classes follow a Bernoulli distribution:

\[
p(x|\theta) = \theta^n (1 - \theta)^{(1-n)} \text{ for } x \in \{0, 1\}.
\]

Part a. Write and submit a program main1_2.m that computes and returns the estimates of the parameters of the Naive Bayes model using the training set pima_train.txt. The parameters include priors on classes, 16 class-conditionals \((8 \times 2 = 16)\), one for every input component and class label. To fit exponential distributions use Matlab function expfit; to fit normal distributions use function normfit (see also Matlab help).
Part b. List parameters found by your program in the report.

Problem 2.3. Classification with the Naive Bayes model

Once the parameters of the Naive Bayes model are learned (estimated), the decision about the class for a specific input $x$ can be made by designing the appropriate discriminant functions. Typically, there are based on class posteriors, thus a classification problem boils down to the problem of comparison of posteriors of classes for $x$. These are computed through the Bayes rule:

$$p(y = 1|x) = \frac{\left[\prod_{i=1}^{d} p(x_i|y = 1)\right] p(y = 1)}{\left[\prod_{i=1}^{d} p(x_i|y = 0)\right] p(y = 0) + \left[\prod_{i=1}^{d} p(x_i|y = 1)\right] p(y = 1)}.$$

Note that in order to make the best posterior choice it is sufficient to compare the following discriminant functions based on log posteriors:

$$g_0(x) = \left[\sum_{i=1}^{d} \log p(x_i|y = 0)\right] + \log p(y = 0) \quad (1)$$

$$g_1(x) = \left[\sum_{i=1}^{d} \log p(x_i|y = 1)\right] + \log p(y = 1) \quad (2)$$

Part a. Write and submit a program `main1_3.m` that:
• Calls a function \textit{predict\_NB} that predicts class labels for inputs based on class posterior. The discriminant functions you need to use here are given in expressions 1 and 2 and use parameters obtained in Problem 2.2.

• Uses \textit{predict\_NB} to compute the misclassification error of the Naive Bayes classifier on both training and test datasets. Report the errors.

• Calculates and reports a \textit{confusion matrix} for the test and training sets.

\textbf{Part b.} In your report include:

• Training and test misclassification errors.

• Confusion matrices for the train and test sets.

• Sensitivity and specificity of the model on the test set.

\textbf{Part c.} Compare results for the mean misclassification errors for Naive Bayes classifier with the results of the logistic regression model from previous assignment (assignment 4).

\textbf{Problem 2. Support vector machines}

Support vector machines represent yet another technique one can apply to the problem of binary classification. The idea is to find the hyperplane that separates the examples in two classes the best. The best hyperplane is defined in terms of the maximum margin. The learning problem reduces as usually to optimization, in this case, a quadratic optimization problem.

There is a number of implementations of SVM algorithms with better or worse running time performances. Here we use a Matlab code implementing SVM solver for the linear decision boundary proposed by O.L. Mangasarian and D. Musicant. The paper describing this method can be downloaded electronically at: http://www.ai.mit.edu/projects/jmlr/papers/volume1/mangasarian01a/html/. The SVM solver is in files \textit{svml.m} and \textit{svml\_itsol.m} that can be downloaded from the course web page. \textit{svml\_itsol.m} is a slightly modified version of the original program by O.L. Mangasarian and D. Musicant. To run it you call \textit{svml.m} that takes care of converting outputs from 0,1 class labels to -1,1 (!!!) and sets other parameters of the Lagrangian SVM.

\textbf{Part a.} Use the \textit{svml} code to learn the weights $w$ and $b$ (bias) of the linear model on the training set. Assume the cost for crossing the boundary (a parameter of the \textit{svml} procedure) is 1.
Part b. Write and submit a function \textit{apply.svml}(x, w, b) that takes an input vector \(x\), and parameters \(w\), and \(b\) of the linear SVM and outputs the class decision (use 0 and 1) for the input \(x\). Briefly, the class is 1 if \(w^T x + b \geq 0\), and 0 otherwise.

Part c. Use function \textit{apply.svml}(x, w, b) to calculate the confusion matrices and other stats for both the training and testing data. More specifically, please report the confusion matrix, misclassification error, sensitivity and specificity for both the training and testing set.

Part d. Compare the above results to the results for the logistic regression (Assignment 4) and the Naive Bayes (problem 1) models. Which model do you think performed the best?

Optional. If you are interested experimenting with existing SVMs tools including tools supporting non-linear kernels please check out the following software packages: liblinear, libsvm and svmlight. All these can be interfaced with Matlab.

Problem 3. ROC analysis

The ROC analysis let us explore the ability of the classification model to discriminate in between the two classes including possible sensitivity and specificity trade-offs. In the ROC analysis we assume a changing threshold for calling class 1 based on the projection defined by the model. This can be \(P(y = 1| x)\) for the logistic regression and the Naive Bayes, or \(g_1(x) = w^T x + b\) for the SVM.

Part a: Familiarize yourself with the function \textit{perfcurve} in matlab that lets you calculate coordinates of points defining the ROC curve, as well as, the area under the ROC curve (AUROC).

Part b. Use the function \textit{perfcurve} to plot the ROC curves and calculate AUROC on the testing set for the Logistic regression from homework assignment 4, and the Naive Bayes and SVM models from this assignment. All models should be trained on the training set.

Part c. Please include the ROC curves and the AUC statistics in the report. Compare the ROC curves and their AUC statistics. What do you think, which model is better?