Introduction to Machine Learning - Homework 2

Prof. Alan Yuille

Spring 2014

Due on Tuesday 13/May. 2014. Hand in papers in class.

Some problems involve programming and data. MATLAB is highly recommended because machine learning algorithms are supported in MATLAB. Octave is free, and has almost the same syntax as MATLAB. The R language also has good machine learning package. For the support vector machine, LIBSVM is an excellent library and is freely available online. You can also use other languages for the homework problems.

Question 1. Regression

Suppose $\vec{x} = (x_1, x_2)$ and we want to perform linear regression in two-dimensions using the function:

$$f(\vec{x}) = \vec{\lambda} \cdot \vec{\phi}(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \tag{1}$$

What is the MLE formulation for this gaussian linear regression? Give a solution for the coefficient $\vec{\lambda}$ and variance σ^2 using a dataset $X = \{(x_1^i, x_2^i), y^i\}_{i=1}^N$.

Now consider polynomial regression in one-dimension with $X = \{x_i, y_i\}_{i=1}^N$. Use file "P1_train.txt" as training data and "P1_test.txt" as test data. The first column of the text file is x_i and the second column is y_i. The n-th order polynomial regression

can be written as

$$f(x) = w_0 + w_1 x + w_2 x^2 + ... + w_n x^n$$

Try first-order and third-order polynomial regression. Give solutions (weights, variances) and the mean-square-error (MSE).

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2.$$

Which method (first or third order regression) gives a better fit to the training data?

Use these methods, learnt on the training data, to make predictions on the test data and plot your results. Again compute the MSE. Which method gives a better fit on the test data? Briefly discuss your results in terms of memorization and generalization. (You can use command "dlmread" to read the text file if you use MATLAB.)

Question 2. AdaBoost.

What is a weak classifier? What is a strong classifier? How does AdaBoost select the weak classifiers? What criterion does AdaBoost minimize? How does this relate to the empirical risk?

Apply AdaBoost to the one-dimensional problem where the data lies on the x-axis. There is one positive example at $x = 0$ and two negative examples at $x = \pm 1$. There are three weak classifiers are $h_1(x) = 1,\ x > 1/2, \&\ h_1(x) = -1,\ x < 1/2$, $h_2(x) = 1,\ x > -1/2, \&\ h_2(x) = -1,\ x < -1/2$, and $h_3(x) = 1,\ \forall x$. Show that this data can be classified correctly by a strong classifier which uses only three weak classifiers. Calculate the first two iterations of AdaBoost for this problem. Are they sufficient to classify the data correctly?

Hint: Weak classifers should always be at least better than 50% prediction accuracy. If not, change the sign.

Question 3. Linear Support Vecctor Machine (SVM)

What is the margin? How does SVM deal with non-separable data? What is the primal formulation of SVM? How does the SVM objective function relate to the empirical risk?

Now consider a binary classification task where the feature is two-dimensional. Use "P3.txt" as data to train a linear SVM. Give the solution (coefficient and bias). Draw the decision boundary.

Learn a Gaussian distribution for each class data: $p(\vec{x}|y = 1)$ with parameter $(\vec{\mu}_1, \sigma_1^2)$ and $p(\vec{x}|y = -1)$ with parameter $(\vec{\mu}_{-1}, \sigma_{-1}^2)$. Give the MLE estimation of parameters. Derive the decision boundary using the log-likelihood ratio. Draw the boundary and compare it with the boundary given by SVM.

Redo the above process for both methods using the other dadaset "P3_outlier.txt". Give the solutions. Draw the boundaries and compare them.

Based on the results from two datasets, which method is better? Why?

Note: The last column of the text file is the class label, and the remaining columns are the features. In the SVM experiment, set $\lambda = 1$. The variable λ is the parameter to balance the margin term and risk term (see handout). In some software, e.g., LIBSVM, C is used instead of λ.

Question 4. Kernel Support Vector Machine (SVM)

What is a kernel? How does it relate to feature vectors?

Consider a binary classification task where the features are high-dimensional. Use "P4_train.txt" as data to train a kernel SVM, and test on "P4_test.txt" to obtain the prediction accuracy. Compare the result of Quadratic kernel (see below) and Radial Basis Function (RBF) kernel.

Use the K-nn classifier to make prediction. Compare the result with those by kernel SVM.

The Quadratic kernel is

$$K(x, y) = (\gamma x \cdot y + \alpha)^2$$

The RBF kernel is

$$K(x, y) = \exp(-\gamma ||x - y||^2)$$

For kernel SVM: Search over $\gamma = (0.01, 0.1, 1)$ and $\lambda = (0.1, 1, 10)$ to obtain the best choice in terms of accuracy using cross-validation. Always set $\alpha = 0$.

For K-nn: Search over $K = (1, 5, 10)$.

Question 5. Primal Dual Quadratic Optimization

The primal formulation is given by:

$$L_p(\vec{a}, b, \{z_i\}; \{\alpha_i, \mu_i\}) = (1/2)|\vec{a}|^2 + \gamma \sum_{i=1}^{m} z_i - \sum_{i=1}^{m} \alpha_i \{y_i(\vec{a} \cdot \vec{x}_i + b) - (1 - z_i)\} - \sum_{i=1}^{m} \mu_i z_i. \tag{2}$$

Explain the meaning of all the terms and variables in this equation. What constraints do the variables satisfy? Calculate the form of the solution $\vec{a}$ by minimizing L_p with respect to $\vec{a}$. What are the support vectors? Obtain the dual formulation by eliminating $\vec{a}, b, \{z_i\}$ from L_p.