Machine Learning Extra Resit Exam

With Answers

12 August 2013

Question	*Points*
Short answers	15
Decision Trees	20
Instance-Based Learning	15
Combining Multiple Learners	20
Comparison of ML algorithms	15
Total	**85**

The exam is **open book**: you can use Alpaydin's *Introduction to Machine Learning* (as a print of the PDF or as a proper book) as well as the lecture slides and any notes you've taken. You can use a calculator. The questions add up to *80 points* in total.

Good luck!

Answers in bold italics.

Questions

1. **Short answers, no justification required** *(3 Points for each question)*

   (a) (**True** or **False**): The standard ID3 decision tree algorithm builds separate trees for each training sample.
   False

   (b) (**True** or **False**): Kernel functions are often used in connection with support vector machines because they allow for non-linear decision surfaces.
   True

   (c) (**True** or **False**): The underlying assumption when using a naive Bayes classifier is that all variables are sampled from a normal distribution
   False

   (d) (**True** or **False**): Clustering algorithms such as EM use unlabelled training data.
   True

   (e) (**True** or **False**): Neural networks can only be used for classification tasks.
   False

2. **Decision Trees**

Consider the following set of training examples:

Classification	*a1*	*a2*
+	T	T
+	T	T
-	T	F
+	F	F
-	F	T
-	F	T

(a) *(5 points)* What is the entropy of this collection of training examples with respect to the classification?

(b) *(5 points)* What is the information gain of attribute *a2* relative to these training examples?

(c) *(5 points)* Does pruning a decision tree such as that produced by the basic (ID3) algorithm increase or decrease performance on the training set? What about the performance on the test set?

(d) *(5 points)* Consider a classification problem with four binary attributes, A, B, C and D, in which the classification is positive if either A=B=0 or C=D=0 and negative otherwise. Draw a decision tree for this problem.

3. **Combining Multiple Learners**

(a) *(7 points)* When combining multiple learners, it is preferable to find a set of diverse learners that differ in their decisions, so that they complement each other. Taking as an example a classification task, list and explain at least two different approaches you can follow to obtain diverse classifiers from your dataset.

Diverse learners can be trained over the same data by using different algorithms (e.g., neural networks, naive Bayes classifiers, ...), or if using the same algorithm, by tuning it in different ways (such as by setting a different number of hidden layer neurons in the neural network. Additionally, instead of training learners over a dataset's complete set of input features, a partitioning of the input features can be used, to train different learners over the different subsets (e.g., a neural network handling a data instance's continuous variables, while a C4.5 decision tree handles the categorical variables). Another possibility is to train the different base-learners over different subsets of the training set (which is the approach followed by bagging and boosting). An answer of just "bagging or boosting" here would be wrong (or more wrong than right). In part because they implement only one of the approaches above (dataset partitioning), but mainly because they are specific implementations of an approach, and the question asks for the approaches themselves.

(b) *(6 points)* Consider a system with multiple sensors producing different kinds of data for each event it logs: a photo, a sound file, as well as

numerical values of temperature, pressure and humidity. You want to train a classifier that will tell you for future logged events whether it is raining or not. By itself, no single kind of data will give you a classifier that is accurate enough, but also no single learner at your disposal will enable you to consider all of an event's input data together. Name two methods to combine the outputs from different classifiers working in parallel over the different kinds of data and explain briefly how they work.

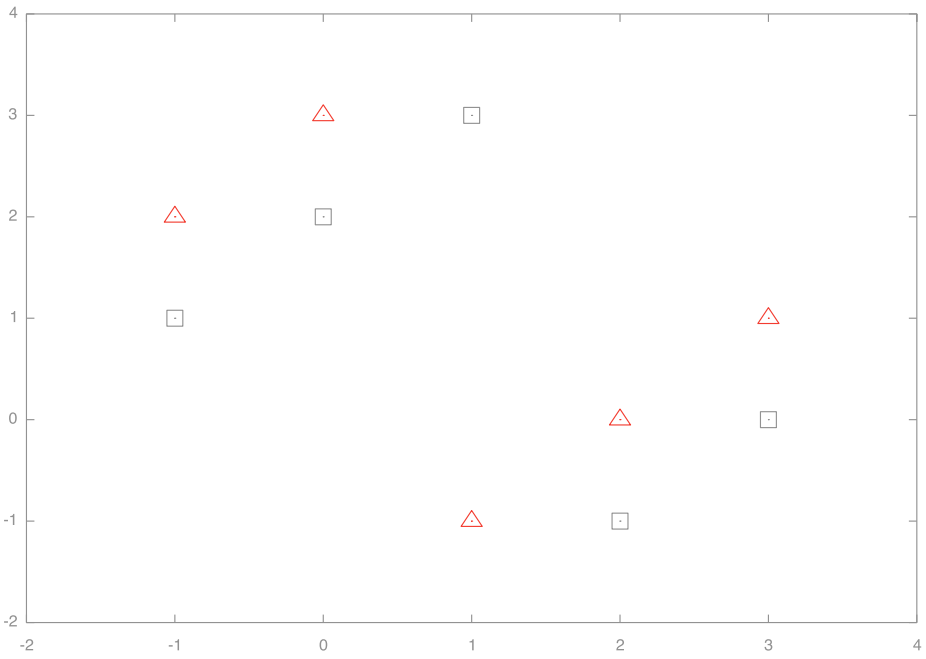
Possible approaches to aggregate the decisions of multiple classifiers include voting, mixture of experts, and stacking.

(c) *(7 points)* You are working with a dataset where you get comparable classification accuracies from both a decision tree inducer and a K-nearest neighbours classifier. Given that there's a fair amount of noise in the data, you then consider using Bagging to improve your classification. Which of the two classifiers would you expect to benefit the most from Bagging? Motivate your answer.

The decision tree inducer would benefit the most. Bagging works best with unstable algorithms, as they generate the most diverse learners from the different training sets given to them by Bagging. (bonus) Furthermore, decision tree inducers can easily be parametrized to make them even more unstable (for instance, by switching pruning off), and thus of even greater benefit in such a setup.

4. **Instance-Based Learning**

Consider K-Nearest Neighbour (k-NN) using Euclidean distance on the following data set (each point belongs to one of two classes: △ or □)



(a) *(5 points)* What is the leave one out cross validation error when using 1-NN?

Every point is misclassified: $10/10 = 1$***.***

(b) *(5 points)* Which of the following values of k leads to the minimum number of leave one out cross validation errors: 3, 5 or 9? What is the error for that k?

All 3 values of k ***misclassify all of the points and have the same classification error as part 1:*** $10/10 = 1$***.***

(c) *(5 points)* Someone suggests a distance weighting function that halves the distance between points on the diagonal. For instance, the distance between $(0, 0)$ and $(1, 1)$ is no longer $\sqrt{2} \approx 1.41$, but now is $\frac{1}{2}\sqrt{2} \approx 0.7$. Other distances remain unaltered. With this distance weighting, what is the leave one out cross validation error when using 1-NN?

All points are now correctly classified: the error is 0***.***

5. **Comparison of ML algorithms**

(a) *(3 points)* Explain in your own words why you cannot use training set error as a reliable estimate of algorithm performance.

(b) *(4 points)* Someone has developed a classifier that correctly classifies 8554 out of 10000 test examples. Give an estimate of the true error of this classifier, and a 95% confidence interval around that estimate (note: $Z_{95} = 1.96$). You may give your confidence interval in the form of an expression.

The answer is the following (where : $\frac{1446}{10000} \pm 1.96 \times \sqrt{\frac{0.1446 \times (1-0.1446)}{10000}} = 0.1446 \pm 0.0035$

(c) *(8 points)* Using a flow-chart or pseudo-code, describe the N-fold cross validation procedure for selecting a classifier algorithm (for instance, choosing between SVM and Decision Trees).