6.867 Machine Learning

Problem Set 4 Solutions

November 9, 2004

Problem 1: Feature Selection

1. Based on the provided data we obtain the following entropies (all logarithms are to the base 2):

$$\begin{array}{ll} H(Y) & 0.9321 \\ H(Y|x=0) & 0 \\ H(Y|x=1) & 0.65 \\ H(Y|x=2) & 1 \\ H(Y|X) & 0.7739 \end{array}$$

Suppose now we add a new edge. Let $H_1(Y)$ be the entropy at the root, and $H_2(Y)$ be the entropy of the label distribution at the added node. Depending on which edge we add, the entropies take the following values:

	$H_1(Y)$	$H_2(Y)$	root count	leaf count	$H(Y\|X)$
edge = 0	0.9024	0	22	1	0.8614
edge = 1	0.9940	0.65	11	12	0.8145
edge = 2	0.7793	1	13	10	0.8753

Note that the entropy at the root changed from $H(Y)$ to $H_1(Y)$ because some of the samples at the root are now taken to the added node.

Because originaly the conditional entropy was $H(Y) = 0.9321$, the reduction in conditional entropy due to the addition of an edge is:

$$\begin{array}{ll} \text{edge} = 0 & 0.0707 \\ \text{edge} = 1 & 0.1176 \\ \text{edge} = 2 & 0.0568 \end{array}$$

Therefore the best edge to add is 1.

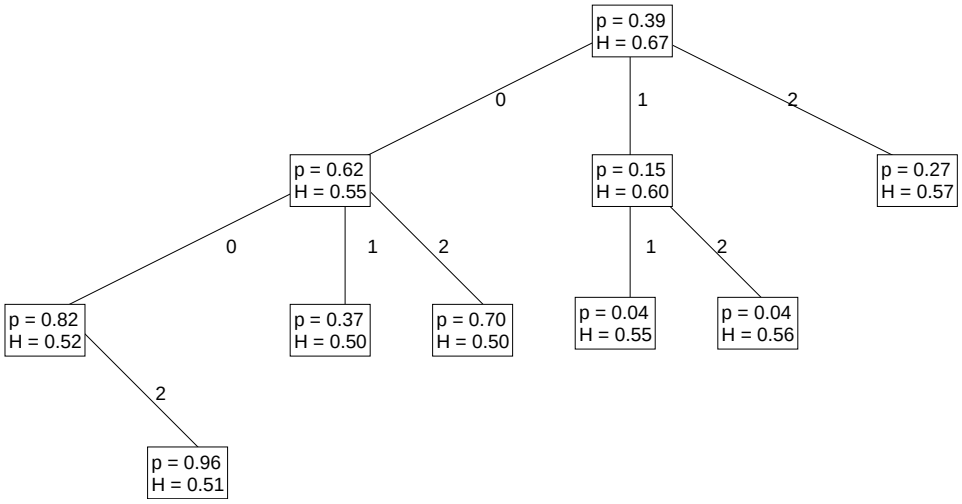

Figure 1: PST tree for Problem 1 Part 2

2. The error rates in the order in which the nodes are added are the following:

$$0.3846, 0.3846, 0.3846, 0.3846, 0.3846, 0.2912, 0.2912, 0.2912, 0.2912, 0.2474$$

The actual PST tree is show in Figure 1.

3. $y = 1$ is most probable when we follow the path $0 \rightarrow 0 \rightarrow 2$. The corresponding rule is:

> If a word starts with two consonants the next letter is very likely to be a vowel.

The path that makes $y = 1$ least likely is either $1 \rightarrow 1$ or $1 \rightarrow 2$ (you can choose one). The corresponding rules are:

> It is very unlikely for a vowel to follow two vowels.

or

> It is very unlikely for a word that begins with a vowel to continue with a vowel.

These rules are consistent with common intuition about English.

4. No, we cannot extend a node past testing for "feature = 2".

Once we the test for "feature = 2" succeeds, all following features will also be equal to 2, because we are past the beginning of the word. Thus all datapoints under this node have the exact feature representation, and we have no basis on which to split them into two groups to further reduce the conditional entropy.

Problem 2

1. **Solution:**

$$\begin{aligned}
\epsilon_m &= \frac{1}{2} - \frac{1}{2}\sum_i \tilde{W}_i^{(m-1)} y_i h(x_i;\hat{\theta}_m) && (1)\\
&= \frac{1}{2}\left[\sum_i \tilde{W}_i^{m-1} - \left(\sum_{i:y_i=h(x_i;\hat{\theta}_m)} \tilde{W}_i^{(m-1)} - \sum_{i:y_i\neq h(x_i;\hat{\theta}_m)} \tilde{W}_i^{(m-1)}\right)\right] && (2)\\
&= \frac{1}{2}\left[\tilde{W}_+^{(m-1)} + \tilde{W}_-^{(m-1)} - \left(\tilde{W}_+^{(m-1)} - \tilde{W}_-^{(m-1)}\right)\right] && (3)\\
&= \tilde{W}_-^{(m-1)} && (4)
\end{aligned}$$

■

2. **Solution:** As stated in the problem,

$$Z_m(\alpha_m) = \tilde{W}_+^{(m-1)}\exp(-\alpha_m) + \tilde{W}_-^{(m-1)}\exp(\alpha_m) \quad (5)$$

Differentiating w.r.t. α_m,

$$\begin{aligned}
\frac{\partial Z_m(\alpha_m)}{\partial \alpha_m} &= -\alpha_m\left(\tilde{W}_+^{(m-1)}\exp(-\alpha_m)\right) + \alpha_m\left(\tilde{W}_-^{(m-1)}\exp(\alpha_m)\right) && (6)\\
&= 0 && (7)
\end{aligned}$$

Therefore,

$$\alpha_m = \frac{1}{2}\log\left(\frac{\tilde{W}_+^{(m-1)}}{\tilde{W}_-^{(m-1)}}\right) \quad (8)$$

Using the result from the previous section,

$$\alpha_m = \frac{1}{2}\log\left(\frac{1-\epsilon_m}{\epsilon_m}\right) \quad (9)$$

■

3. Minimum value of $Z_m(\alpha_m)$ is:

$$\begin{aligned}
Z_{m,\min} &= \tilde{W}_+^{(m-1)}\exp\left(\frac{1}{2}\log\left(\frac{\tilde{W}_-^{(m-1)}}{\tilde{W}_+^{(m-1)}}\right)\right) + \tilde{W}_-^{(m-1)}\exp\left(\frac{1}{2}\log\left(\frac{\tilde{W}_+^{(m-1)}}{\tilde{W}_-^{(m-1)}}\right)\right)\\
&= \tilde{W}_+^{(m-1)}\sqrt{\frac{\tilde{W}_-^{(m-1)}}{\tilde{W}_+^{(m-1)}}} + \tilde{W}_-^{(m-1)}\sqrt{\frac{\tilde{W}_+^{(m-1)}}{\tilde{W}_-^{(m-1)}}}\\
&= 2\sqrt{\tilde{W}_+^{(m-1)}\tilde{W}_-^{(m-1)}}\\
&= 2\sqrt{(1-\epsilon_m)\epsilon_m}
\end{aligned}$$

Using induction on $L(h_m) = L(h_{m-1})Z_m$, we get

$$L(h_m) = \prod_{k=1}^{m} Z_k = \prod_{k=1}^{m} 2\sqrt{(1-\epsilon_k)\epsilon_k} \tag{10}$$

Since we know that the error is $\leq L(h_m)$, we get the required result.

4. **Solution:**

   (a) `alpha = 0.5*log((1-stump.werr)/stump.werr);`

   (b) The following code snippet calculates the minimum voting margin among the training examples:

```
for k=1:num_iter
  [hhtrain,summtrain]=eval_boost(model(1:k),data.xtrain(:,1:2));
  votemargintrain(k)=(min(hhtrain.*data.ytrain))/summtrain;
end
figure;
plot(votemargintrain);
xlabel('Number of boosting iterations');
ylabel('Voting margin train');
title('Voting margin as a function of the number of iterations');
```

   See Figures 2 and 3.

   The test error decreases initially and then remains constant. However, the minimum voting margin on the training examples increases. It crosses zero (at which point all training samples are correctly classified), and converges to a value slightly less that 0.2. With more test samples, further reduction in (percentage) test error might have been observed.

   (c) See Figure 4.

■

5. **Solution:** No. We are jointly optimising the votes here (in contrast to the greedy optimisation in AdaBoost). Also, the votes need not be positive here. ■

Problem 3

1. **Solution:** The VC dimension of the given class of classifiers is 3.

   Three points in general position (*i.e.* 3 non-collinear points) can clearly be shattered by this set of classifiers. However, no set of 4 points can be shattered. To see this, we consider two cases (ignoring the cases of 3 or more collinear points, which can clearly not be shattered):

– The convex hull of the 4 points is a triangle, with one point lying strictly inside this convex hull: in this case, labelling the points at the vertices of the triangle as +1 and the interior point as -1 is not possible.

– The convex hull of the 4 points is a quadrilateral: in this case, one of the two labellings of non-adjacent vertices—both +1, remaining vertices -1, or both -1, remaining vertices +1—is not possible.

■

2. **Solution:**

(a) VC dimension is 2. This can be seen as follows.

Given two decision stump classifiers $h_1(x)$ and $h_2(x)$, the classifier obtained as a convex combination is given by $\text{sign}(\alpha_1 h_1(x)+\alpha_2 h_2(x))$. As stated in the problem set, a single vertical decision stump (and hence also a convex combination of two vertical decision stumps) can shatter 2 points in $\Re^2$.

However, no set of 3 points can be shattered. For the purpose of labelling points using vertical decision stumps, we need only consider the horizontal coordinates of the points. Let these be x_1, x_2 and x_3. Further, let sign(0) be equal to +1. In this case, the labelling $x_1 = -1$, $x_2 = +1$, $x_3 = -1$ is not possible. Changing the definition of sign(0) does not help, as then the case $x_1 = +1$, $x_2 = -1$, $x_3 = +1$ is not possible.

(b) VC dimension is 3.

Consider 3 points that form an equilateral triangle, one of whose sides is parallel to the horizontal axis. Any required labelling of these 3 points can be obtained by using either a single horizontal or a single vertical decision stump. Thus, the set of convex combinations of a horizontal and a vertical decision stump (in particular, the subset where one of the two weights in the combination is unity and the other zero) can shatter 3 points.

No set of 4 points can be shattered by the given set of classifiers. Let us assume that sign(0) is equal to +1; it is easy to show that the opposite assumption leads to equivalent results. Then the possible decision regions include axis-aligned half spaces (with either label) or axis-aligned quadrants (with label -1). The latter case arises when the two stumps have equal weights in the convex combination.

Consider two cases (ignoring the cases of 3 or more collinear points, which can clearly not be shattered):

∗ The convex hull of the 4 points is a triangle, with one point lying strictly inside this convex hull: in this case, labelling the points at the vertices of the triangle as -1 and the interior point as +1 is not possible. This is because the interior point cannot lie in a half-space that does not contain any of the other three points.

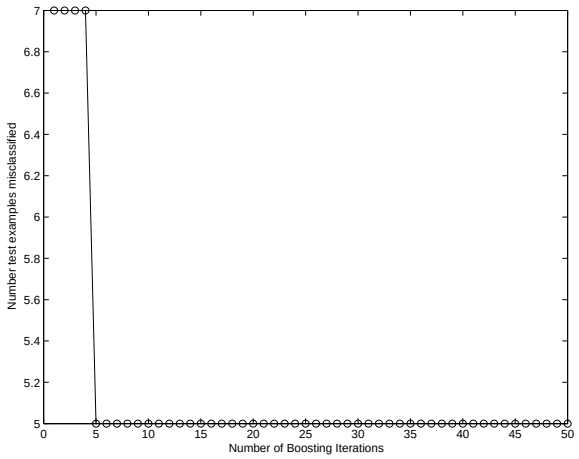

Figure 2: Test error for AdaBoost classifier

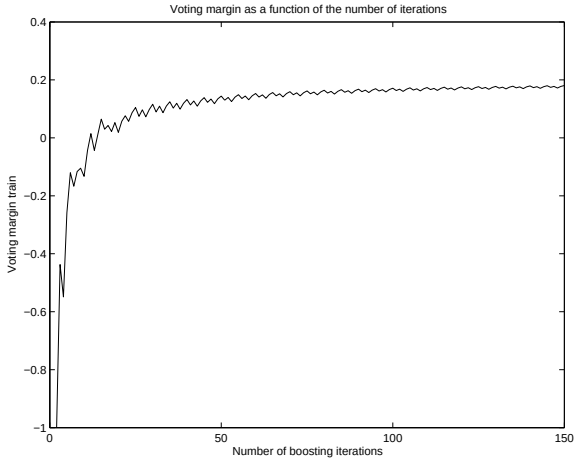

Figure 3: Minimum voting margin over training examples for AdaBoost classifier

∗ The convex hull of the 4 points is a quadrilateral: in this case, one of the two labellings of non-adjacent vertices—both +1, remaining vertices -1, or both -1, remaining vertices +1—is not possible.

■

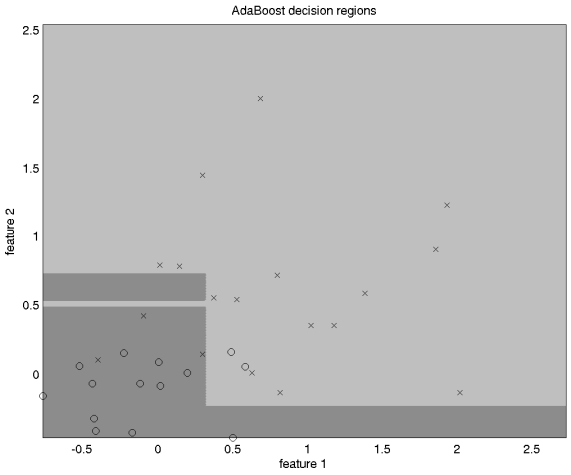

Figure 4: Decision regions for AdaBoost classifier