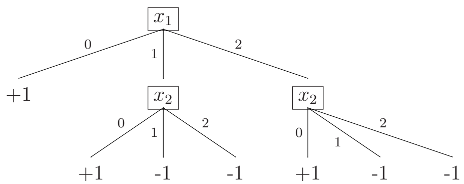
Answers Resit Exam Machine Learning

February 18, 2008

1. (a)



   (b) h_2 is inconsistent with the third example in Table 1 and is therefore eliminated by the algorithm. h_1 and h_3 are consistent with all training data in Table 1. Both h_1 and h_3 classify the new instance as -1. Therefore the LIST-THEN-ELIMINATE algorithm also classifies the new instance as -1.

   (c) For example,

$$h_4(\mathbf{x}) = \begin{cases} +1 & \text{if } x_1 = 0, \\ +1 & \text{if } x_1 = 2 \text{ and } x_2 = 1, \\ -1 & \text{otherwise.} \end{cases}$$

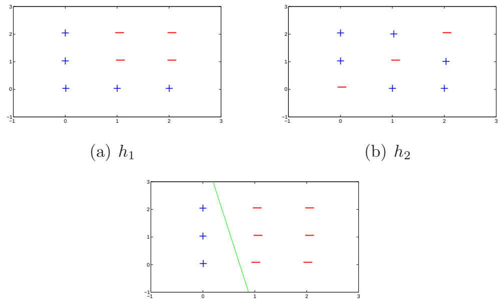
   (d) Only h_3 can be implemented by a perceptron with suitable weights. The other two hypotheses give classifications that are not linearly separable. See Figure 1.

   (e) For example: $C(h_1) = 0$, $C(h_2) = 10$, $C(h_3) = 11$. (See also slide 5 of `mlslides12-part2.pdf`.)

2. Naive Bayes would classify $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ as -1, because:

$$\begin{aligned} P(y=-1)P(x_1 = 1 \mid y=-1)P(x_2 = 2 \mid y = -1) &= \frac{1}{2} \cdot \frac{1}{2} \cdot 1 \\ > \frac{1}{2} \cdot 0 \cdot \frac{1}{2} &= P(y=+1)P(x_1 = 1 \mid y = +1)P(x_2 = 2 \mid y = +1) \end{aligned}$$

3. (a) No, the tree will probably overfit data D and therefore have higher accuracy on D than on new data. (See also Figure 3.6 in Mitchell.)



(a) h_1 (b) h_2 (c) h_3

Figure 1: Graphical illustration of hypotheses in $\mathcal{H}$. Only h_3 has a linear decision boundary.

Another correct answer would be that ID3 grows the tree until it is consistent with all the training data. Unless this tree is completely correct and there is no noise in the data, this tree will make mistakes if we use it to classify data in a new test set. Therefore we cannot expect the tree to have approximately the same accuracy on the new data as on data D.

   (b) No pruning will occur, because ID3 stops growing the tree when it makes no errors on the training data. Therefore any pruning would introduce errors and hence decrease the accuracy on data D. Therefore if D is used to decide which nodes to prune, no nodes will be pruned.

4. (a)

k	1	3	5
Instance 1	W	B	B
Instance 2	W	W	B

   (b) This will multiply the Euclidean distance between any two points by this number. Thus the image will be scaled, but all relative distances remain the same and all examples keep the same neighbours. This does not influence the algorithm.

   (c) This will be relatively easy, because the target function assigns the same label to large regions of the feature space: There is one large White region and one large Black region. Only the instances with features close to the border between the Black and White regions ($x_1 + x_2$ close to 100) will be hard to learn for the algorithm.

5. Maximum likelihood parameter estimation will select P_1:

$$P_1(D) = 0.3^4 \cdot 0.7^4 = \left(\frac{21}{100}\right)^4 > \left(\frac{16}{100}\right)^4 = 0.8^4 \cdot 0.2^4 = P_2(D)$$

Let π be a prior on θ such that $\pi(\theta = 1) = x$ and $\pi(\theta = 2) = 1 - x$. Then MAP will select P_2 if

$$\begin{aligned} P_1(D)\pi(\theta = 1) &< P_2(D)\pi(\theta = 2) \\ \left(\frac{21}{100}\right)^4 x &< \left(\frac{16}{100}\right)^4 (1 - x) \\ \left(\frac{21}{16}\right)^4 x &< 1 - x \\ \left(1 + \left(\frac{21}{16}\right)^4\right) x &< 1 \\ x &< 1/\left(1 + \left(\frac{21}{16}\right)^4\right). \end{aligned}$$

Thus MAP will select a different hypothesis from maximum likelihood for any prior π such that $\pi(\theta = 1) < 1/\left(1 + \left(\frac{21}{16}\right)^4\right) \approx 0.252$.