Department of Mathematics and Computer Science
University of Southern Denmark, Odense

March 15, 2013

Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 14, Spring 2013

Exercise 1
Do exercises 1, 4, 5 from Exam 2010.

Solution

$$J_b = \sum_{j=1}^{m} w_j^b I\{h(\mathbf{x}_j;\theta) \neq y_j\}$$

For A $J_b = 0.06$, for B $J_b = 0.05$ for C $J_b = 0.5$. Hence B is the best model.

Solution The question is unclear since $h(\cdot)$ is not defined. Using the definition from AdaBoost: the prediction is computed by sign $\left(\sum_{b=1}^{B} \alpha_b h_b(\mathbf{x})\right)$. We can use:

$$h(\mathbf{x}) = \text{sign}\left(\alpha_A h_A(\mathbf{x}) + \alpha_C h_c(\mathbf{x})\right)$$

the values of α are given by:

$$\epsilon_b = \frac{\sum_{i=1}^{m} w_i^b I\left\{h_b(\mathbf{x}^i) \neq y^i\right\}}{\sum_i w_i^b} \qquad \alpha_b = \ln \frac{1-\epsilon_b}{\epsilon_b}$$

from which we get $\alpha_A = 2.7$ and $\alpha_C = 0$. Hence the prediction will be the same as for A and wrong in c and d.
Alternatively we assume:

$$h(\mathbf{x}) = \text{sign}\left(h_A(\mathbf{x}) + h_c(\mathbf{x})\right)$$

and get wrong predictions for b, c, d.

Solution We represent an observation of X_1 by a binary vector $\mathbf{x}_1$ with $\sum_{k=1}^{5} x_{1k} = 1$. Further, let $p(X_1 = x_{1k}) = \theta_{1k}$ with $\sum_{k=1}^{5} \theta_{1k} = 1$. Then, the distribution of X_1 is a generalization of the Bernoulli distribution:

$$p(X_1 = x_k|\boldsymbol{\theta}_1) = \prod_{k=1}^{5} \theta_{1k}^{x_{1k}} \tag{1}$$

Similarly for $X_2|X_1$ we represent an observation of X_2 by a binary vector $\mathbf{x}_{2k}$ with $\sum_{l=1}^{3} x_{2kl} = 1$ and $p(X_2 = x_{2kl}|X_1 = x_{1k}) = \theta_{2kl}$ with $\sum_{l=1}^{3} \theta_{2kl} = 1$.

$$p(X_2 = x_l|X_1 = x_k, \boldsymbol{\theta}_{2k}) = \prod_{l=1}^{3} \theta_{2kl}^{x_{2kl}} \tag{2}$$

Solution $P(X_2 = \texttt{linkern}|X_1 = \texttt{nn}, \mathcal{D}) = \prod_{j=1}^{5} \theta_{2kl}^{x_{2kl}^j}$. To estimate θ_{2kl} we write the joint likelihood and use factorization

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{5} P(X_2^j = \mathbf{x}_2, X_1^j = \mathbf{x}_1) \tag{3}$$
$$= \prod_{j=1}^{5} P(X_2^j = \mathbf{x}_2 | X_1^j = \mathbf{x}_1) P(X_1^j = \mathbf{x}) \tag{4}$$
$$= \prod_{j=1}^{5}\prod_{k=1}^{5}\prod_{l=1}^{3} \theta_{2kl}^{x_{2kl}^j} \theta_{1k}^{x_{1k}^j} \tag{5}$$

Then we maximize in $\boldsymbol{\theta}$ with the additional constraint that $\sum_l \theta_{2kl} = 1$, which we Lagrange relax:

$$L'(\theta) = \ell(\theta) + \lambda \sum_l (\theta_{2kl} - 1) \tag{6}$$
$$= \log \prod_{j=1}^{5}\prod_{k=1}^{5}\prod_{l=1}^{3} \theta_{2kl}^{x_{2kl}^j} \theta_{1k}^{x_{1k}^j} + \lambda \sum_l (\theta_{2kl} - 1) \tag{7}$$
$$= \sum_{j=1}^{5} x_{2kl}^j \log \theta_{2kl} + \sum_{j=1}^{5} x_{1k}^j \log \theta_{1k} + \lambda \sum_l (\theta_{2kl} - 1) \tag{8}$$

$$\frac{\partial L(\theta)}{\partial \theta_{2kl}} = \sum_{j=1}^{5} x_{2kl}^j \frac{1}{\theta_{2kl}} + \lambda \theta_{2kl} = 0 \tag{9}$$

Solution This corresponds to $P(X_2 = \texttt{nearest\_insertion}, X_1 = \texttt{nn} | \mathcal{D})$ that we can estimate as in the previous point giving 0.
The problem is that with max likelihood we are overfitting. Laplace smoothing could help.

Solution

$$p(X_1 = x_k) = \theta_{1k} \tag{10}$$
$$p(\boldsymbol{\theta}_1 | \boldsymbol{\alpha}_1) = \text{Dirichlet}(\boldsymbol{\theta}_{1k} | \boldsymbol{\alpha}_1) = \frac{\Gamma(\alpha_{10})}{\Gamma(\alpha_{11}) \cdots \Gamma(\alpha_{15})} \prod_{k=1}^{K} \theta_k^{\alpha_{1k}-1} \tag{11}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. Similarly,

$$p(X_2 = x_l | X_1 = x_k) = \theta_{2kl} \tag{12}$$
$$p(\boldsymbol{\theta}_{2k} | \boldsymbol{\alpha}_{2k}) = \text{Dirichlet}(\boldsymbol{\theta}_{2k} | \boldsymbol{\alpha}_{2k}) \tag{13}$$

Solution A uniform initial local prior means that all hyperparameters are equal to 1. In general, for $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]^T \sim \text{Dir}(\boldsymbol{\alpha})$

$$E[\theta_i] = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$$

In our case, we first calculate the posterior probability

$$\begin{aligned}
p(\theta_{2kl}|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta_{2kl})p(\theta_{2kl})}{p(\mathcal{D})} && (14)\\
&= \mathrm{Dir}(\theta_{2kl}|\boldsymbol{\alpha}_{2k} + \mathbf{m}_{2k}) && (15)\\
&= \frac{\Gamma(\alpha_{2k0} + N)}{\Gamma(\alpha_{2k1} + m_{2k1}) \cdots \Gamma(\alpha_{2k3} + m_{2k3})} \prod_{l=1}^{3} \theta_{2kl}^{\alpha_{2kl}+m_{2kl}-1} && (16)
\end{aligned}$$

where we have denoted $\mathbf{m}_{2k} = [m_{2k1}, \ldots, m_{2k3}]^T$, the number of observations of x_{2kl} and $N = \sum_l m_{2kl}$. Then we calculate $P(X_2 = x_{1l}|X_1 = x_{1k}, \mathcal{D})$ by marginalizing over the parameter θ_{2kl}. This will give us the expected value of θ_{2kl} with respect to its Dirichelet distribution:

$$\begin{aligned}
P(X_2 = x_{1l}|X_1 = x_{1k}, \mathcal{D}) &= \int_{-\infty}^{+\infty} P(X_2 = x_{2l}|X_1 = x_{1k}, \theta_{2kl}\mathcal{D})p(\theta_{2kl}|\mathcal{D})d\theta_{2kl} && (17)\\
&= E_{p(\theta_{2kl}|\mathcal{D})}[\theta_{2kl}|\mathcal{D}] && (18)\\
&= \frac{\alpha_{2kl} + m_{2kl}}{\sum_l \alpha_{2kl} + N} && (19)\\
&= \frac{1+2}{3+2} = \frac{3}{5} = 0.6 && (20)
\end{aligned}$$

Solution

We need to compute the joint probability distribution $P(X_1, X_2)$

$$\begin{aligned}
P(X_1 = x_{1k}, X_2 = x_{2l}|\mathcal{D}) &= P(X_2 = x_{2l}|X_1 = x_{1k}, \mathcal{D})P(X_1 = x_{1k}|\mathcal{D}) && (21)\\
&= E_{p(\theta_{2kl}|\mathcal{D})}[\theta_{2kl}|\mathcal{D}]E_{p(\theta_{1k}|\mathcal{D})}[\theta_{1k}|\mathcal{D}] && (22)\\
&= \frac{\alpha_{2kl} + m_{2kl}}{\sum_l \alpha_{2kl} + N_2} \frac{\alpha_{1k} + m_{1k}}{\sum_k \alpha_{1k} + N_1} && (23)
\end{aligned}$$

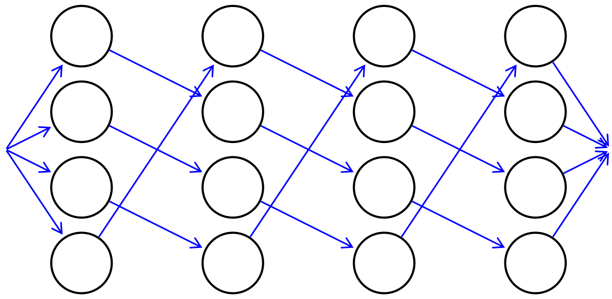
Thus,

$$\begin{aligned}
p(\texttt{arbitrary_insertion-linkern}) &= \frac{1+3}{3+3} \cdot \frac{1+3}{5+5} = \frac{4}{6}\frac{4}{10} = 0.266 && (24)\\
p(\texttt{nn-linkern}) &= \frac{1+2}{3+2} \cdot \frac{1+2}{5+5} = \frac{3}{5}\frac{3}{10} = 0.180 && (25)\\
p(\texttt{arbitrary_insertion-none}) &= \frac{1+0}{3+3} \cdot \frac{1+3}{5+5} = \frac{1}{6}\frac{4}{10} = 0.066 && (26)\\
p(\texttt{farthest_insertion-linkern}) &= \frac{1+0}{3+0} \cdot \frac{1+0}{5+5} = \frac{1}{3}\frac{1}{10} = 0.033 && (27)\\
&\vdots && (28)
\end{aligned}$$

Solution

$$\begin{aligned}
p(X_2 = x_{2l}|\mathcal{D}) &= \sum_k P(X_2 = x_{2l}|X_1 = x_{1k}, \mathcal{D})P(X_1 = x_{1k}|\mathcal{D}) && (29)\\
&= \sum_k E_{p(\theta_{2kl}|\mathcal{D})}[\theta_{2kl}|\mathcal{D}]E_{p(\theta_{1k}|\mathcal{D})}[\theta_{1k}|\mathcal{D}] && (30)\\
&= \sum_k \frac{\alpha_{2kl} + m_{2kl}}{\sum_l \alpha_{2kl} + N_2} \frac{\alpha_{1k} + m_{1k}}{\sum_k \alpha_{1k} + N_1} && (31)
\end{aligned}$$

Solution



Solution The likely states we can get to at $t = 3$ are $x_3 = 2$ and $x_3 = 3$. The other state sequences will have probability at most 10^{-9}. The mean μ_3 is closer to the observation point thus favoring the state sequence that ends up in $x_3 = 3$. This is $x_1 = 1, x_2 = 2, x_3 = 3$.

Solution Roughly, the max probability to these points (could be derived efficiently via max sum algorithm) is:

$$p(y_3|X3 = 3)p(X3 = 3|X2 = 2)p(X_2 = 2|X_1 = 1)p(X_1 = 1) = p_3 \cdot 1 \cdot 1 \cdot 0.5$$

$$p(y_3|X3 = 4)p(X3 = 4|X2 = 3)p(X_2 = 3|X_1 = 2)p(X_1 = 2) = p_4 \cdot 1 \cdot 1 \cdot 10^{-9}$$

hence if σ^2 becomes such that $p_3 \cdot 0.5 < p_4 \cdot 10^{-9}$ then the state sequence that ends up in $x_3 = 4$ ($x_1 = 2, x_2 = 3, x_4 = 4$) will become the most likely state sequence.

Exercise 2 – Tree based methods

Consider a data set comprising 400 data points from class $\mathcal{C}_1$ and 400 data points from class $\mathcal{C}_2$. Suppose that a tree model A splits these into (300,100) assigned to the first leaf node (predicting $\mathcal{C}_1$ and (100,300) assigned to the second leaf node (predicting $\mathcal{C}_2$, where (n, m) denotes that n points come from class $\mathcal{C}_1$ and m points come from class $\mathcal{C}_2$. Similarly, suppose that a second tree model B splits them into (200,400) and (200,0), respectively. Evaluate the misclassification rates for the two trees and show that they are equal. Similarly, evaluate the pruning criterion for the cross-entropy case for the two trees.

Exercise 3 – Tree based methods

You are given the following data points: Negative: (-1, -1) (2, 1) (2, -1); Positive: (-2, 1) (-1, 1) (1,-1). The points are depicted in Figure 1.

1. Construct a decision tree using the greedy recursive bi-partitioning algorithm based on information gain described in class. Use both criteria the Gini index and the entropy. In the search for the split threshold θ discretize the continue scale of the two features and consider only values in $\{-1.5, 0, 1.5\}$ for f_1 and $\{0\}$ for f_2. Represent graphically the tree constructed and draw the decision boundaries in the Figure 1. Table 1 might be useful for some computations

2. Use the tree to predict the outcome for the new point (1,1).

Exercise 4 – Nearest Neighbor

x	y	$-(x/y) \cdot \log(x/y)$	x	y	$-(x/y) \cdot \log(x/y)$
1	2	0.50	1	5	0.46
1	3	0.53	2	5	0.53
2	3	0.39	3	5	0.44
1	4	0.50	4	5	0.26
3	4	0.31			

Table 1: Numerical values for the computation of information gains.

1. Draw the decision boundaries for 1-Nearest Neighbor on the Figure 1. Make it accurate enough so that it is possible to tell whether the integer-valued coordinate points in the diagram are on the boundary or, if not, which region they are in.
2. What class does 1-NN predict for the new point: (1, 1).
3. What class does 3-NN predict for the new point: (1, 0).

Exercise 5 – Practical

Analyze by means of classification tree the data on spam email from the UCI repository. Use `rpart` from the `rpart` package and the `ctree` from the `party` package.

Exercise 6 – PCA

Using the `iris` data readily available in R use principle component analysis to identify two components and plot the data in these components. Can you classify the data at this stage?

Exercise 7 – Probability and Independence

A joint probability table for the binary variables A, B, and C is given below.

A / B	b_1	b_2
a_1	(0.006, 0.054)	(0.048, 0.432)
a_2	(0.014, 0.126)	(0.032, 0.288)

Table 2: Joint probability distribution $P(A, B, C)$

- Calculate $P(B, C)$ and $P(B)$.
- Are A and C independent given B? (Remember to report the justification of your answer.)

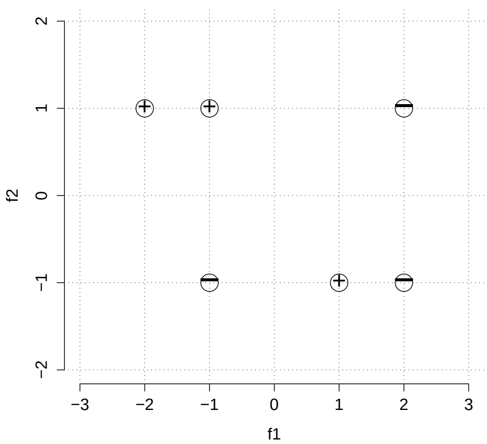


Figure 1: The data points for classification.