C19 Machine Learning

A. Zisserman, Hilary Term 2015

1. Given the following training data for a $\{0, 1\}$ binary classifier:

$$(x_1 = 0.8; y_1 = 1); \ \ (x_2 = 0.4; y_2 = 0); \ \ (x_3 = 0.6; y_3 = 1)$$

Determine the output of a K Nearest Neighbour (K-NN) classifier for all points on the interval $0 \leq x \leq 1$ using

   (a) 1-NN

   (b) 3-NN

2. A regressor algorithm is defined using the mean of the K Nearest Neighbours of a test point. Determine the ouput on the interval $0 \leq x \leq 1$ using the training data in question (1) for K=2.

3. Two students are working on a machine-learning approach to spam detection. Each student has their own set of 100 labeled emails, 90% of which are used for training and 10% for validating the model. Student A runs a K-NN classification algorithm and reports 80% accuracy on her validation set. Student B experiments with over 100 different learning algorithms, training each one on his training set, and recording the accuracy on the validation set. His best formulation achieves 90% accuracy. Whose algorithm would you pick for protecting a corporate network from spam? Why?

4. For a linear SVM, show that the vector $\mathbf{w}$ in the primal cost function can be expressed as $\sum_i^N \alpha_i y_i \mathbf{x}_i$, where $\{\mathbf{x}_i, y_i\}$ are the training data. (Hint, start by expressing $\mathbf{w} = \sum_i^N \alpha_i y_i \mathbf{x}_i + \mathbf{w}_\perp$, where $\mathbf{w}_\perp$ is the subspace orthogonal to $\mathbf{x}_i \forall i$).

5. Suppose that a linear SVM is learnt in the dual form from the training data $\{\mathbf{x}_j, y_j\}$ so that the support vectors $\mathbf{x}_s$ and α_s are known.

   (a) How can the weight vector and bias, $\mathbf{w}$, b, of the primal form be obtained?

   (b) Why is it an advantage to use the classifier in the primal form?

6. (a) Determine the mapping $\phi(\mathbf{x})$ such that the kernel

$$k(\mathbf{x}, \mathbf{z}) = (c + \mathbf{x}^\top \mathbf{z})^2 = \phi(\mathbf{x})^\top \phi(\mathbf{z})$$

where $\mathbf{x} = (x_1, x_2)^\top$ and $\mathbf{z} = (z_1, z_2)^\top$.

(b) Show, by a sketch, that an XOR is not linearly separable, but that after the mapping $\phi(\mathbf{x})$ with $c = 0$ it is linearly separable

7. (a) Show that if the SVM cost function is written as

$$\mathcal{C}(\mathbf{w}) = \frac{1}{N} \sum_i^N \left(\frac{\lambda}{2} ||\mathbf{w}||^2 + \max\left(0, 1 - y_i f(\mathbf{x}_i)\right) \right)$$

where $f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i$, then using using steepest descent optimization, $\mathbf{w}_{t+1}$ may be learnt from $\mathbf{w}_t$ by cycling through the data with the following update rule

$$\begin{aligned} \mathbf{w}_{t+1} \ &\leftarrow \ (1 - \eta\lambda)\mathbf{w}_t + \eta y_i \mathbf{x}_i \quad \text{if } y_i \mathbf{w}^\top \mathbf{x}_i < 1 \\ &\leftarrow \ (1 - \eta\lambda)\mathbf{w}_t \qquad\qquad \text{otherwise} \end{aligned}$$

where η is the learning rate.

   (b) Contrast the SVM update rule with that of the perceptron

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \text{sign}(\mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i$$

(NB, there is no update if the point is correctly classified, i.e. $y_i \mathbf{w}^\top \mathbf{x}_i > 0$). What are the differences, and how do they influence the margin?

(c) The perceptron learning rule can be derived as steepest descent optimization of a loss function. What is the loss function?

8. A K-class discriminant is obtained by training K linear classifiers of the form

$$f_k(\mathbf{x}) = \mathbf{w}_k{}^\top \mathbf{x} + b_k$$

and assigning a point to class C_k if $f_k(\mathbf{x}) > f_j(\mathbf{x})$ for all $j \neq k$.

(a) Write the equation of the hyperplane separating class j and k.

(b) If $\mathbf{x}_\mathrm{A}$ and $\mathbf{x}_\mathrm{B}$ are both in the decision region R_j (i.e. classified as class j), then show that any point on the line

$$\mathbf{x} = \lambda \mathbf{x}_\mathrm{A} + (1 - \lambda)\mathbf{x}_\mathrm{B}$$

where $0 \leq \lambda \leq 1$, is also classified as class j.

9. During training of a four class decision tree, 100 samples of each class arrive at the node, and the output of a set of three possible node tests are shown in the table:

class	test 1		test 2		test 3	
	L	R	L	R	L	R
A	100	0	100	0	40	60
B	0	100	100	0	60	40
C	0	100	0	100	40	60
D	0	100	0	100	60	40

where L and R refer to the number of each class that are sent to the left and right child nodes.

(a) Compute the information gain for each node test given by

$$I = H(S) - \sum_{i \in L,R} \frac{S_i}{S} H(S_i)$$

where S is the number of samples arriving at the node, S_i the number sent to each child, and the entropy of the set s is given by $H(s) = -\sum_j p_j \log p_j$ with p_j the probability of class j in the set.

(b) Using the information gain, decide which test should be chosen. Sketch the child probability distributions to check your result.

(c) Does the base of the log matter in making this choice?

10. A student uses the regression function

$$f(x, \mathbf{w}) = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \ldots + w_M\phi_M(x) = \mathbf{w}^\top \Phi(x)$$

(where x is a scalar and f a scalar valued function) for two possible data sources:

(a) A periodic source which oscillates with a known period p.

(b) A polynomial of second degree.

What are suitable basis functions for each of these sources? Can the student save time and design a single set of basis functions $\phi_i(x)$ that will allow him/her to model observations from either source?