CM311: Model Answers

Steve Gunn

9th November 2001

Question 1

a) **Describe what is meant by the terms *training set*, *validation set* and *testing set*.**

(3 marks)

Training set is the data used for learning. Validation set is the data used for determining hyperparameters. Testing set is the data used for estimating generalisation performance.

b) **Describe what is meant by the terms *classification*, *regression* and *density estimation*.**

(3 marks)

Classification involves learning the distinction between two or more classes of data. Regression involves learning a mapping from the input space to a real variable. Density estimation involves learning a distribution from a finite set of exmples.

c) **Describe a useful real-world application of machine learning.**

(4 marks)

E.g. Handwriting recognition. Candidate should discuss data collection, preprocessing, learning algorithm and factors limiting overall performance.

Question 2

a) **Explain what is meant by the term over-parameterisation. State how it is removed from the hyperplane, $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b=0$, in the linear Support Vector Machine formulation to produce a canonical hyperplane.**

(3 marks)

Over-parameterisation refers to case when the parameters in an equation are not independent, and hence two different sets of parameters can describe an identical solution. The over-parameterisation in $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b = 0$ is removed by adding the constraint $\min_i \left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right| = 1$ where $\boldsymbol{x}_i$ are the input space coordinates of the training examples.

b) **State the condition for separability of the two-class data-set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ with this canonical hyperplane.**

(3 marks)

$y_i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] \geq 1, \quad i = 1, \ldots, n.$

c) **Describe the maximum margin principle and show that the resulting optimisation problem is given by the Lagrangian,**

$$\Phi(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^n \alpha_i \left(y_i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] - 1\right), \quad \alpha_i \geq 0.$$

(8 marks)

The maximum margin principle states: "The set of vectors, $\boldsymbol{x}_i$, is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vector to the hyperplane is maximal."

The distance $d(\boldsymbol{w}, b; \boldsymbol{x})$ of a point $\boldsymbol{x}$ from the hyperplane $(\boldsymbol{w}, b)$ is,

$$d(\boldsymbol{w}, b; \boldsymbol{x}) = \frac{\left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|}.$$

The optimal hyperplane is given by maximising the margin, ρ, subject to the separability constraints of b). The margin is given by,

$$\begin{aligned}
\rho(\boldsymbol{w}, b) &= \min_{\boldsymbol{x}_i : y_i = -1} d(\boldsymbol{w}, b; \boldsymbol{x}_i) + \min_{x_i : y_i = 1} d(\boldsymbol{w}, b; \boldsymbol{x}_i) \\
&= \min_{\boldsymbol{x}_i : y_i = -1} \frac{\left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x}_i : y_i = 1} \frac{\left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|} \\
&= \frac{1}{\|\boldsymbol{w}\|}\left(\min_{\boldsymbol{x}_i : y_i = -1} \left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right| + \min_{\boldsymbol{x}_i : y_i = 1} \left|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right|\right) \\
&= \frac{2}{\|\boldsymbol{w}\|}
\end{aligned}$$

Hence the hyperplane that optimally separates the data is the one that minimises

$$\Phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2.$$

subject to $y^i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] \geq 1$ and hence the Lagrangian is,

$$\Phi(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^n \alpha_i \left(y_i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] - 1\right), \quad \alpha_i \geq 0.$$

d) **Solve the Lagrangian problem,** $\max_{\boldsymbol{\alpha}} \left(\min_{\boldsymbol{w},b} \Phi(\boldsymbol{w}, b, \boldsymbol{\alpha})\right)$**, to show that the solution for the Lagrange multipliers can be written as a quadratic program,**

$$\min_{\boldsymbol{\alpha}} \tfrac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}} H \boldsymbol{\alpha} + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\alpha},$$

subject to the constraints,

$$\alpha_i \geq 0, \quad \textstyle\sum_{j=1}^{n} \alpha_j y_j = 0.$$

(8 marks)

The minimum with respect to $\boldsymbol{w}$ and b of the Lagrangian, Φ, is given by,

$$\frac{\partial \Phi}{\partial b} = 0 \;\Rightarrow\; \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial \Phi}{\partial \boldsymbol{w}} = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i.$$

Hence, by substitution and rearrangement the dual problem is,

$$\max_{\boldsymbol{\alpha}} -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j + \sum_{k=1}^{n} \alpha_k,$$

and hence the solution to the problem is given by,

$$\min_{\boldsymbol{\alpha}} \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j - \sum_{k=1}^{n} \alpha_k,$$

with constraints,

$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \alpha_j y_j = 0.$$

This is equivalent to the quadratic program formulation by noting $H_{i,j} = y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j$ and $c_i = -1$.

e) **What are the Support Vectors and how do these relate to the Lagrange multipliers?**

(3 marks)

The support vectors are the elements in the data-set which lie on the margin. Each data point has an associated Lagrange multiplier and only the support vectors have non-zero Lagrange multipliers. (Strictly speaking the Support Vectors can also lie on one side of the margin in non-separable cases, but the candidate can gain full marks without noting this.)

Question 3

a) **Describe the K Nearest Neighbour method.**

(8 marks)

K Nearest Neighbour is a classification algorithm whereby a voting scheme is used to determine the K points in the training set which lie nearest to the test point. Candidate should draw diagram in two dimensional input space to illustrate this. Various possibilities for measuring distance, but common one, unless otherwise specified, is Euclidean distance. Typically K is chosen to be odd in a two class problem since this will typically force a majority vote (unless two distances are equal). K controls overfitting of the technique - large K smoother boundary, smaller K more sensitive to noise in the data. Algorithm typically performs well in comparison to state-of-the-art methods such as SVMs. It is always possible to find the global solution and whilst no explicit model is built the algorithm defines an underlying model.

b) **Describe the methods of holdout validation, cross-validation and bootstrap validation.**

(9 marks)

Validation is the method of using some of the data to determine hyperparameters in a learning algorithm, e.g. K in the K nearest neighbour algorithm. Holdout validation splits the original data set into two parts - one for training and one for validation. However, when the dataset is small this may leave a small amount of data for learning which is bad. Hence, the need for cross-validation and bootstrap validation. Cross validation splits the data into n sets and uses each of the n sets in turn for validation, whilst using the remaining data for training. In the extreme case we have leave-one-out cross validation whereby one data point is used for validation and the rest for training. The hyperparameter is determined by taking the best mean performance on all the validation sets. Bootstrap validation allows an infinite number of training and validation sets to be obtained by sampling with replacement from the original data set to obtain as many sets for training and validation as required. In classification problems the data is usually stratified such that the class priors are reflected in each data set.

c) **Given the following 2-class dataset, use stratified 4-fold cross validation to estimate the optimal value for K, in the K nearest neighbour algorithm, and discuss the result.**

(16 marks)

x_1	x_2	y
2.5	1.2	A
5.3	2.3	A
4.2	1.1	A
5.3	2.3	A
1.2	2.7	B
4.2	3.4	B
2.3	1.3	B
1.2	2.7	B

Number the data points as x^i for $i = 1 \ldots 8$. The training, $\mathcal{T}_i$, and validation sets, $\mathcal{V}_i$ are given by $\mathcal{T}_1 = \{x^2, x^3, x^4, x^6, x^7, x^8\}, \mathcal{V}_1 = \{x^1, x^5\}$ $\mathcal{T}_2 = \{x^1, x^3, x^4, x^5, x^7, x^8\}, \mathcal{V}_2 = \{x^2, x^6\}$ $\mathcal{T}_3 = \{x^1, x^2, x^4, x^5, x^6, x^8\}, \mathcal{V}_3 = \{x^3, x^7\}$ $\mathcal{T}_4 = \{x^1, x^2, x^3, x^5, x^6, x^7\}, \mathcal{V}_4 = \{x^4, x^8\}$. (There are a couple of other possibilities here that will also receive full marks.)

The distances are given by

	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x^1	3.0083	1.7029	3.0083	1.9849	2.7803	0.2236	1.9849
x^2		1.6279	0	4.1195	1.5556	3.1623	4.1195
x^3			1.6279	3.4000	2.3000	1.9105	3.4000
x^4				4.1195	1.5556	3.1623	4.1195
x^5					3.0806	1.7804	0
x^6						2.8320	3.0806
x^7							1.7804

Build table for validation error against K. Only need to consider $K \in \{1, 2, 3\}$ since training set size has only 3 examples per class. (Use 0.5 error when classes are equal) Hence $K = 3$ is optimal. (Note the candidate may get a different answer if they choose one of the other possible combinations for the CV sets - they will obviously get full marks if they do this correctly.)

K	$\mathcal{V}_1$	$\mathcal{V}_2$	$\mathcal{V}_3$	$\mathcal{V}_4$	Total Error
1	1	1	1	0	3
2	0.5	0.5	1	0.5	2.5
3	1	1	0	0	2

The data is overlapping and hence either noisy or contains insufficient features for precise classification. Consequently, a higher value of K is appropriate to regularise the model.