19s1: COMP9417 Machine Learning and Data Mining

Lectures: Kernel Methods
Topic: Questions from lecture topics
Version: with answers
Last revision: Thu Jul 11 2019

Introduction

Some questions and exercises from the course lectures covering "Kernel Methods", focusing on Vapnik's Support Vector Machine (SVM) for classification tasks.

Question 1 Review the development of the Perceptron training algorithm on slides 6–12 of the lecture "Neural Learning". Now compare this to the algorithm for Perceptron training *in dual form* introduced on slides 19–20 of the lecture "Kernel Methods". Basically, the two algorithms differ only around lines 6–7. Provide an explanation of how the dual version of the algorithm relates to the original.

Answer
For the purposes of this answer, we ignore the learning rate η, or equivalently assume that $\eta = 1$. In the original Perceptron training algorithm we try to classify the current example and check for an error, which occurs if $y_i \mathbf{w} \cdot \mathbf{x}_i \leq 0$. When this happens, we update the weight vector $\mathbf{w}$ by adding to it $\eta y_i \mathbf{x}_i$, i.e., $y_i \mathbf{x}_i$.
Thinking about this, clearly the weight vector that is learned will have been updated zero or more times for *every* example that has been misclassified. Denoting this number as α_i for example $\mathbf{x}_i$, the weight vector can be expressed as

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

On this *dual* view, we are learning instance weights α_i rather than feature weights w_j. An instance $\mathbf{x}$ is classified as

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} \right)$$

During training, the only information needed about the training data is all pairwise dot products: the n-by-n matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ containing these dot products is called the *Gram matrix.*

Question 2 The Support Vector Machine is a essentially an approach to learning linear classifiers, but uses a alternative objective function to methods we looked at before, namely *maximising the margin.* Learning algorithms for this problem typically use quadratic optimization solvers, but it is possible to derive the solution manually for a small number of support vectors.

Here is a toy data set of three examples shown as the matrix $\mathbf{X}$, of which the first two are classified as positive and the third as negative, shown as the vector $\mathbf{y}$. Start by constructing the *Gram matrix* for this data, incorporating the class labels, i.e., form the matrix $\mathbf{X}'\mathbf{X}'^{\mathrm{T}}$. Then solve to find the support vectors, their Lagrange multipliers α, then determine the weight vector $\mathbf{w}$, threshold t and the margin m.

$$\mathbf{X} = \begin{pmatrix} 1 & 3 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} +1 \\ +1 \\ -1 \end{pmatrix} \qquad \mathbf{X}' = \begin{pmatrix} 1 & 3 \\ 2 & 1 \\ 0 & -1 \end{pmatrix}$$

Background To find a maximum margin classifier requires finding a solution for $\mathbf{w}$, t and margin m. For this we can use the following steps (refer to slides 30–35 from the "Kernel Methods" lecture):

1. set up Gram matrix for labelled data
2. set up expression to be minimised
3. take partial derivatives
4. set to zero and solve for each multiplier
5. solve for $\mathbf{w}$
6. solve for t
7. solve for m

Answer

Step 1. The Gram matrix is

$$\mathbf{X}'\mathbf{X}'^{\mathrm{T}} = \begin{pmatrix} 10 & 5 & -3 \\ 5 & 5 & -1 \\ -3 & -1 & 1 \end{pmatrix}$$

Step 2. The dual optimisation problem is thus

$$\operatorname*{argmax}_{\alpha_1,\alpha_2,\alpha_3} -\frac{1}{2}\left(10\alpha_1^2 + 10\alpha_1\alpha_2 - 6\alpha_1\alpha_3 + 5\alpha_2^2 - 2\alpha_2\alpha_3 + \alpha_3^2\right) + \alpha_1 + \alpha_2 + \alpha_3$$

subject to $\alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0$ and $\alpha_1 + \alpha_2 - \alpha_3 = 0$. From this we obtain the next expression, and its following simplification

$$\begin{aligned} &\operatorname*{argmax}_{\alpha_1,\alpha_2,\alpha_3} -\frac{1}{2}\left(10\alpha_1^2 + 10\alpha_1\alpha_2 - 6\alpha_1(\alpha_1+\alpha_2) + 5\alpha_2^2 - 2\alpha_2(\alpha_1+\alpha_2) + (\alpha_1+\alpha_2)^2\right) + 2\alpha_1 + 2\alpha_2 \\ = &\operatorname*{argmax}_{\alpha_1,\alpha_2,\alpha_3} -\frac{1}{2}\left(5\alpha_1^2 + 4\alpha_1\alpha_2 + 4\alpha_2^2\right) + 2\alpha_1 + 2\alpha_2 \end{aligned}$$

Step 3.

$$\frac{\partial}{\partial\alpha_1} = -5\alpha_1 - 2\alpha_2 + 2$$

$$\frac{\partial}{\partial\alpha_2} = -2\alpha_1 - 4\alpha_2 + 2$$

Step 4. Setting partial derivatives to zero and solving gives $\alpha_1 = \frac{1}{4}$ and $\alpha_2 = \frac{3}{8}$. Also, since $\alpha_3 = \alpha_1 + \alpha_2$ we have $\alpha_3 = \frac{5}{8}$.

Step 5. From slide 30 $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ so

$$\begin{aligned} \mathbf{w} &= \frac{1}{4}\mathbf{x}_1 + \frac{3}{8}\mathbf{x}_2 - \frac{5}{8}\mathbf{x}_3 \\ &= \begin{pmatrix} 1 \\ 1/2 \end{pmatrix} \end{aligned}$$

Step 6. t can be obtained from any support vector, say $\mathbf{x}_3$, since $y_3(\mathbf{w}\cdot\mathbf{x}_3 - t) = 1$; this gives $t = \frac{3}{2}$.

Step 7. Finally, the margin $1/||\mathbf{w}|| = 1/\sqrt{(1^2 + (\frac{1}{2})^2)} = \frac{4}{5}$.

Question 3 You are given 4 examples from a text classification problem where the feature values are simple word counts. Furthermore, all the examples are support vectors, with coefficients $\alpha_i = 1$. Here are the support vectors:

goal	referee	anti	campaign	ban	national	Class
2	0	1	0	4	1	+1
0	3	0	1	0	2	+1
1	0	3	2	0	2	-1
0	2	2	4	0	2	-1

Use the classification rule

$$\hat{y} = \sum_{\mathbf{x}_i \text{ is a support vector}} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x}$$

to classify the example $\mathbf{x} = (1, 1, 0, 0, 3, 3)$. What is the class $\hat{y}$ of this example ? Now try this one and determine its class: $\mathbf{x} = (0, 1, 2, 3, 1, 1)$. (It's probably easier to write a short script to do this.)

Answer

```
Running first ...
x: [2 0 1 0 4 1] qx: 17 qxy: 17 c: 17.0
x: [0 3 0 1 0 2] qx: 9 qxy: 9 c: 9.0
x: [1 0 3 2 0 2] qx: 7 qxy: -7 c: -7.0
x: [0 2 2 4 0 2] qx: 8 qxy: -8 c: -8.0
Classification: 11.0
Running second ...
x: [2 0 1 0 4 1] qx: 7 qxy: 7 c: 7.0
x: [0 3 0 1 0 2] qx: 8 qxy: 8 c: 8.0
x: [1 0 3 2 0 2] qx: 14 qxy: -14 c: -14.0
x: [0 2 2 4 0 2] qx: 20 qxy: -20 c: -20.0
Classification: -19.0
```

Question 4 You are told that the "kernel trick" means that a non-linear mapping can be realised from the original data representation to a new, implicit feature space simply by defining a kernel function on dot products of pairs of instances from the original data. To see why this is so, you take two instances $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$, and take their dot product $\mathbf{x} \cdot \mathbf{y}$ and obtain the answer 7. Clearly, raising this dot product to the power of two will give $(\mathbf{x} \cdot \mathbf{y})^2 = 49$.

Now expand out this expression to show that this is the same answer you would have obtained if you had simply done a set of feature transformations on the original data.

Answer

The expansion is simply (for data instances with exactly two features)

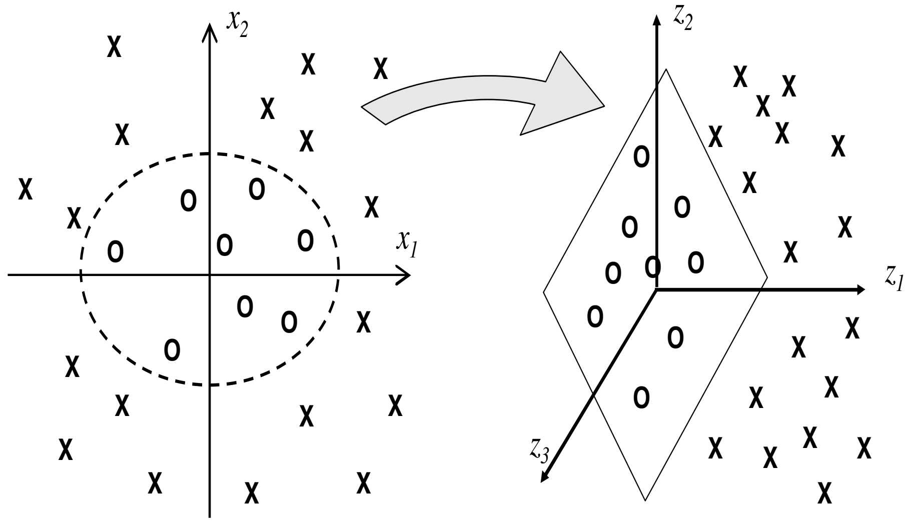
$$\begin{aligned}(\mathbf{x}\cdot\mathbf{y})\times(\mathbf{x}\cdot\mathbf{y}) &= (x_1y_1 + x_2y_2)\times(x_1y_1 + x_2y_2)\\ &= (x_1y_1)^2 + 2(x_1y_1x_2y2) + (x_2y_2)^2\end{aligned}$$

Plugging in our data we get

$$\begin{aligned}&= (1\times 3)^2 + 2(1\times 3\times 2\times 2) + (2\times 2)^2\\ &= 9 + 24 + 16\\ &= 49\end{aligned}$$

Question 5 Generate at least one example from each of the two classes in the original feature space (left of diagram below) and apply the feature transformation (at top of diagram below) to show that the transformation gives rise to a linear separating hyperplane in the new feature space (right of diagram below).

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$$



Answer
Here is one answer: 'X' class: $(5,5)$ maps to $25, 25, 35)$. 'O' class: $(-1,-1)$ maps to $(1, 1, 1.41)$.
