T5

Friday, June 11, 2021 8:55 PM

Tutorial5

STATS5099: Data Mining

Tutorial Sheet 5

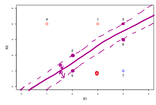
All questions are optional (not the exam-style questions). They are designed to enhance understanding of lecture materials.

Conceptual

1. **Support vectors and margins**

Consider the two-class dataset below:

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue



Left: a two-dimensional dataset with 7 observations; *Right*: scatterplot of the dataset

(a) Sketch the decision boundary that would be found by a linear SVM for this dataset and provide its equation.
Hint: The decision boundary is essentially the maximal margin hyperplane, which should have the form of $w_1x_1 + w_2x_2 + b = 0$.

(b) Describe the classification rule of SVM.

(c) On your sketch, indicate the margin and the support vectors.

(d) Imagine that a new blue data point is added to this dataset at position (3,1). How would this affect the decision boundary if,

- $C = 0$
- $C = 1$
- $C = \infty$

(e) Repeat question (d) with a red data point added at (3, 1).

(a) mid-point between 2 and 3: (2,2), (2,1) → (2, 1.5)
mid-pt between 3 and 6: (3,4), (4,3) → (4, 3.5)
⇒ Set $w_1 = 1$ and solve w_1, b
⇒ $w_1X_1 + X_2 + b = 0$
⇒ $\begin{cases} 2w_1 + 1.5 + b = 0 \\ 4w_1 + 3.5 + b = 0 \end{cases}$ ⇒ $\begin{cases} w_1 = 1 \\ b = -3.5 \end{cases}$

(b) assign observation to red class if $-X_1 + X_2 - 3.5 > 0$

(d) A blue pt added at (3,1) is outside the margin, and it will not affect the decision boundary.

(e) A red pt at (3,1) is on wrong side of decision boundary to satisfy $y_i(w^Tx+b) \geq 1-\xi_i$, we need to include ξ_i
If $C=0$, the slightest change will not be allowed, so the decision boundary will remain the same.
If $C=1$, we expect a right move toward (3,1) in order to account for the penalty associated with misclassifying an obs pt.
If $C=\infty$, the SVM optimization will fail. There is no linear decision boundary, and there is an infinite penalty associated w/ misclassifications

2. **Kernel functions**

Support vector machines are able to produce non-linear decision boundaries by, in a sense, transforming low-dimensional inputs into a high-dimensional space, then performing classification in that high-dimensional space. This usually works because high-dimensional data is much more likely to be linearly separable than low-dimensional data.

As an example, consider the following two-dimensional data points: $\boldsymbol{x}_i = (1, 2)$, $\boldsymbol{x}_j = (2, 4)$. Use the following function to map these two data points into the indicated six-dimensional space (x_1 denotes the first feature, x_2 denotes the second feature):

$$\phi(\boldsymbol{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

(a) $\phi(\boldsymbol{x}_i) =$ $(1, 4, 2\sqrt{2}, \sqrt{2}, 2\sqrt{2}, 1)$

(b) $\phi(\boldsymbol{x}_j) =$ $(4, 16, 8\sqrt{2}, 2\sqrt{2}, 4\sqrt{2}, 1)$

Now calculate the inner product between these two points in the space defined by ϕ:

(c) $\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j) =$ 121

The kernel function corresponding to the projection above is the polynomial kernel of degree 2:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j + 1)^2.$$

Use this kernel to calculate:

(d) $K(\boldsymbol{x}_i, \boldsymbol{x}_j) =$ $(1 \times 2 + 2 \times 4 + 1)^2 = 121$

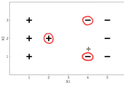
If all went well, your answers for (c) and (d) should be the same. Obviously, using the kernel function required significantly fewer calculations. The difference would be even more dramatic if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ were three-dimensional data points, or if we were to select a higher degree polynomial.

The beauty of non-linear support vector machines is that training and prediction only require the inner products between data points. The data points themselves don't show up anywhere in the calculations. This means that the work you did in (a), (b) and (c) is never actually performed by the SVM. It "cheats" by using the Kernel function instead. This cheating is referred to as the *kernel trick*.

Acknowledgement: Question 2 is designed by Dr Nathan Sprague.

1. Support vectors are the observations that lie closest to the decision boundary. TRUE/FALSE

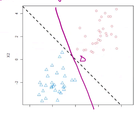
2. Suppose you are using a linear SVM classifier for binary classification problem. Now you have been given the following data in which some points are circled red representing support vectors.



If you remove any one red point from the data, will the decision boundary change? YES/NO

3. The straight line represents the decision boundary of a linear SVM for this data set. TRUE/FALSE

not max margin



$\xi_i = 0$

4. If the data are linearly separable, a linear SVM will return the same parameters $\boldsymbol{w}$ and b regardless of the value of the cost parameter C. TRUE/FALSE

5. When the cost parameter C is set to infinite, which of the following holds true?

(a) The separating hyperplane, if exists, will be the one that completely separates the data.
(b) The soft-margin classifier will separate the data. data is not seperatable ⇒ cannot seperate
(c) Both of the above.
(d) None of the above.

6. The effectiveness of an SVM depends upon: *(select all that apply)*

(a) The distribution of the data (non-parametric)
(b) Selection of kernel

(c) Kernel parameters
(d) Soft margin cost parameter C

7. After training a soft-margin SVM with a linear kernel, you find both training and validation accuracy are low. What will you consider next?
underfitted ⇒ add complexity
(a) Increase the cost parameter C (not care about margin ⇒ better seperate)
(b) Decrease the cost parameter C
(c) Include more features
(d) Use a nonlinear kernel

8. An SVM with a polynomial kernel of degree 2 achieves 100% accuracy training and validation data sets. What if the degree is now changed to 5, assuming all other parameters are fixed?
more complex ⇒ overfitted ⇒ train ↑ ⇒ validation ↓ test ↓
(a) Nothing will change.
(b) Training accuracy becomes lower, but validation accuracy remains 100%.
(c) Training accuracy remains 100%, but validation accuracy becomes lower.
(d) Both training and validation accuracy become lower.

9. Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?
$\gamma \uparrow \Rightarrow$ distance small
⇒ only choose pt close
(a) The model will consider far away points from hyperplane for modelling.
(b) The model will consider only the points close to the hyperplane for modelling.
(c) The model will not be affected by distance of points from hyperplane for modelling.
(d) None of the above.

11. The following is the decision boundary of an SVM with a RBF kernel with Gamma=10. Which is the boundary of Gamma=5000?
model complex ⇒ decision boundary wiggly

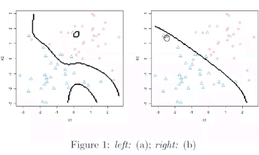


Figure 1: *left:* (a); *right:* (b)