Discussion 5
Machine Learning,Spring 2018

1 KKT Conditions

1.1 Inequality Constraint

Let $\alpha \in \mathbb{R}$ and $a \in \mathbb{R}^n$ with $a \neq 0$. Define the halfspace $H = \{x \in \mathbb{R}^n : a^T x + \alpha \geq 0\}$. Consider the problem of finding the point in H with the smallest Euclidean norm.

(a) Formulate this problem as a constrained optimization problem.

(b) Solve the problem with the help of the KKT conditions. (Hint: you should consider different cases based on if α is negative or nonnegative.)

Solution: (a) This is just the optimization problem

$$\min_{x \in \mathbb{R}^n} x^T x \quad \text{such that } a^T x + \alpha \geq 0.$$

(b) We first start by writing the KKT conditions:

$$\mathcal{L}(x, \lambda) = x^T x - \lambda(a^T x + \alpha),$$

and

$$\nabla_x \mathcal{L}(x, \lambda) = 2x - \lambda a.$$

Hence, the KKT conditions are

$$\begin{aligned} 2x - \lambda a &= 0 \\ a^T x + \alpha &\geq 0 \\ \lambda &\geq 0 \\ \lambda(a^T x + \alpha) &= 0. \end{aligned}$$

First note that if $\alpha \geq 0$, then the 0 vector already satisfies the constraint, and clearly it has minimal norm, so in this case, the optimal solution is 0.

Now assume that $\alpha < 0$. Note that the final KKT condition we wrote above implies that $\lambda = 0$ or $a^T x + \alpha = 0$. If $\lambda = 0$, then the very first KKT condition shows that $0 = 2x - 0 = 2x$, so $x = 0$. But this violates the constraint $a^T x + \alpha \geq 0$, since $a^T 0 + \alpha = \alpha < 0$. Hence, we know that $\lambda \neq 0$. Then $a^T x = -\alpha$. On the other hand, from the first KKT condition, we have that $x = \frac{\lambda}{2} a$, and so by substitution, we have

$$-\alpha = \frac{\lambda}{2} a^T a,$$

which means that

$$\lambda = -\frac{2\alpha}{a^T a}.$$

Plugging this back into the first KKT condition gives

$$x = -\frac{\alpha}{a^T a} a.$$

This is the only point that satisfies the KKT conditions, and from there it is easy to verify that it is the solution.

2 SVM

2.1 Concepts

(a) True or False: The maximum margin hyperplane is only defined by the location of the support vectors, thus other data points can be moved around freely (so long as they remain outside the margin region) without changing the decision boundary. Explain.

Solution True. Decision boundary is determined by the support vectors.

(b) True or False: Suppose there is a data set just containing two data points from different classes. If we fit a SVM, then there is a unique solution for the location of the maximum margin hyperplane. Explain.

Solution Generally True (except the case two points overlop).

2.2 Practice

Recall the soft margin SVM primal problem

$$\min \left(\frac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w} + C \sum_{i=1}^{n} s_i \right), \quad \forall i = 1, \ldots, n, \quad s_i \geq 0, \quad (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) y_i \geq (1 - s_i)$$

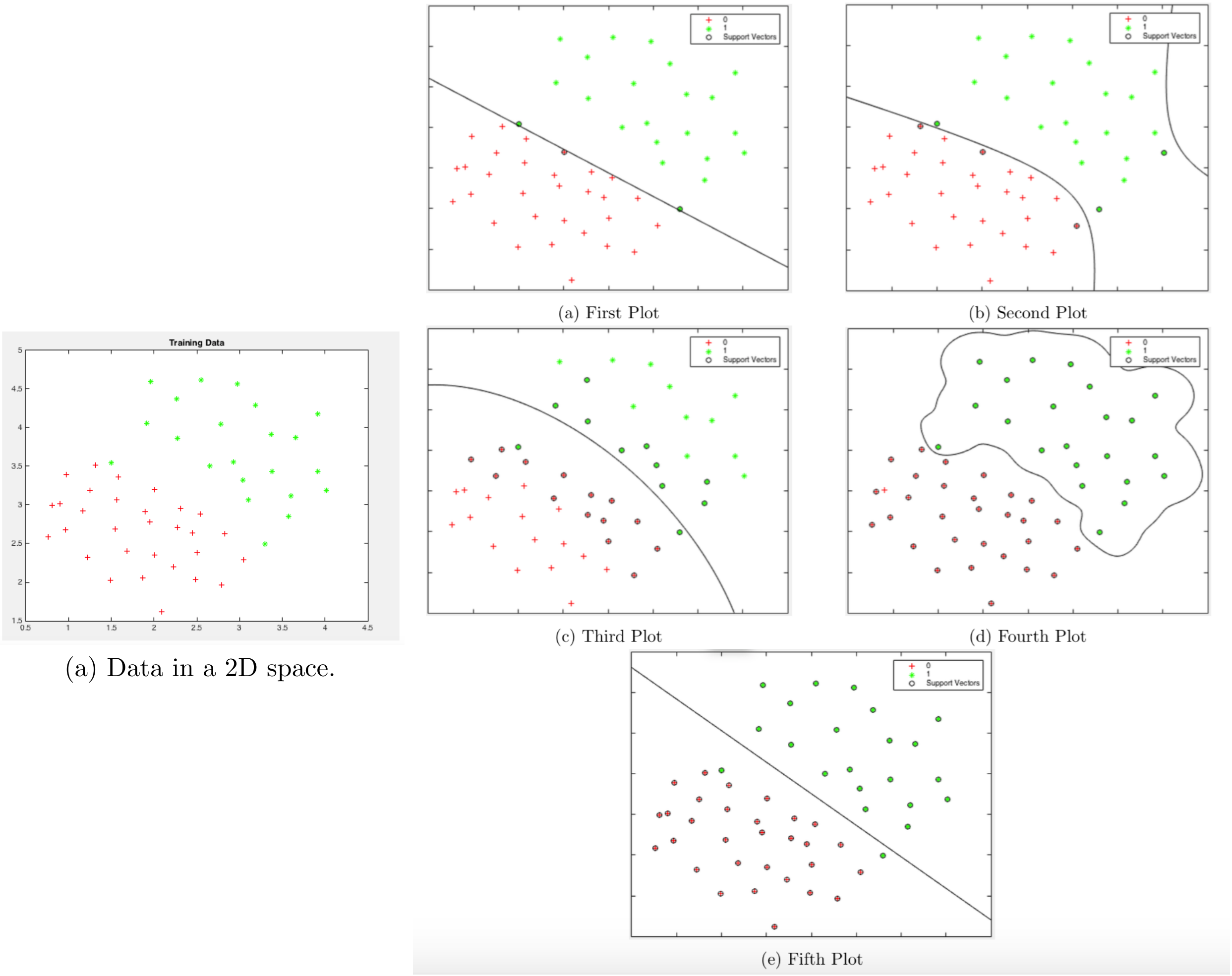
For the hard-margin case, we have $s_i = 0, \forall i$. Varying the C parameter changes the position of the decision boundary. We can obtain the kernel SVM by considering the dual and replacing $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ with $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Consider the following dataset of points with two classes, shown in Fig. 1 (a). We decided to play around with the C parameter and kernel functions and observed how the decision boundary changes. We plotted a couple of pictures (shown in Fig. 1 (b)). However, we were careless and didn't label the plots! Answer the following questions:

(a) Select the plot that best corresponds to a soft-margin linear SVM with $C = 0.01$.
Solution: The first clue is to notice that since we're using a linear SVM, our possible answers are now the first and fifth plots. The second clue is that we have a very low value for C in which case we allow a bit of slack in finding a decision boundary. Hence the fifth plot is more likely for this case.

(b) Select the plot that best corresponds to a soft-margin linear SVM with $C = 100000$.
Solution: Same reasoning as before except we now have a very high C. This means we surely do not want to make any errors in training and so we end up with the very strict decision boundary shown in the first plot.

(c) Suppose we were given information that the data we have might be slightly noisy: a couple of points are moved a short distance away from where they actually should be. Which of the above two classifiers do you think does better in this case?
Solution: We're told that the data might be slightly noisy. Notice how apart from a couple of points, the two classes are fairly far apart. Therefore, to generalize well, one would want to ignore these few points and go for a more general classifier which is the Linear SVM with low C value.



(a) First Plot

(b) Second Plot

(c) Third Plot

(d) Fourth Plot

(e) Fifth Plot

(a) Data in a 2D space.

(b) Testing

Figure 1: Data for Decision Trees.

(d) Say you are allowed to add a new datapoint to the dataset. State briefly that in which cases the decision boundary will not change when using the classifier that you selected for the previous part?
Solution: Only the point that will be a support vector (points closest to the decision boundary), otherwise it wouldn't have an impact.

(e) Select the plot that best corresponds to a soft-margin quadratic kernel SVM with $C = 0.1$.
Solution: We're now using a quadratic kernel. Let's look at the remaining options - the figure with the closed loop is not possible for a quadratic kernel and so we look at the other two option - second and third. Next clue is that we have a low C. Using logic we used before, we can conclude that the corresponding plot is the third plot.

(f) Select the plot that best corresponds to a soft-margin quadratic kernel SVM with $C = 100000$.
Solution: Similar to the previous question but we now have a high C and thus we overfit to the training data. The second plot is the answer here.

(g) Select the plot that best corresponds to a soft-margin gaussian kernel ($\sigma = 0.2$) SVM with $C = 10000$.
Solution: We simply stick with the remaining plot which is the fourth plot. Also, the decision

boundary of Gaussian kernel can be complicated, because it can be written as the sum of polynomials.