CS 189 Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk DIS3

1 Support Vector Machines with Custom Margins

Consider a soft-margin SVM. We are given a training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ and solve the following optimization problem.

$$\text{Choose } \mathbf{w}, \alpha, \xi_i \text{ that minimize} \quad \|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{1}$$

$$\text{such that} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha) \geq 1 - \xi_i \quad \forall i \in [1, n] \tag{2}$$

$$\xi_i \geq 0 \quad \forall i \in [1, n] \tag{3}$$

Today, we are interested in a modified version of the soft-margin SVM where we have a custom margin for each of the n data points. In the standard soft-margin SVM, we pay a penalty of ξ_i for each data point. But we might not want to treat each training point equally, since with prior knowledge, we might know that some data points are more important or more reliable than others (analogous to weighted least-squares regression). We consider a slightly modified optimization problem.

$$\text{Choose } \mathbf{w}, \alpha, \xi_i \text{ that minimize} \quad \|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \phi_i \xi_i \tag{4}$$

$$\text{such that} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha) \geq 1 - \xi_i \quad \forall i \tag{5}$$

$$\xi_i \geq 0 \quad \forall i \tag{6}$$

The only difference is that we have a weighting factor $\phi_i > 0$ for each of the slack variables ξ_i in the objective function. The ϕ_i's are constants based on prior knowledge. This formulation weights each violation ξ_i differently according to the prior knowledge ϕ_i.

(a) For the standard soft-margin SVM, the constrained optimization problem is equal to the following unconstrained optimization problem, known as regularized empirical risk minimization problem with hinge loss.

$$\text{Choose } \mathbf{w}, \alpha \text{ that minimize} \|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\{1 - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha), 0\}. \tag{7}$$

What is the corresponding unconstrained optimization problem for the SVM with custom margins?

Solution: The corresponding unconstrained optimization problem is

$$\text{Choose } \mathbf{w}, \alpha \text{ that minimize} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \phi_i \max\{1 - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha), 0\}. \tag{8}$$

We can see this as follows. Manipulating the first inequality, we have that

$$\xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha) \quad \forall i. \tag{9}$$

Combining this with the second inequality, we have that

$$\xi_i \geq \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha), 0). \tag{10}$$

Since we are minimizing and since we know that $\phi_i > 0$ for all i, we conclude that the constraint must be tight:

$$\xi_i = \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha), 0). \tag{11}$$

The above unconstrained problem then follows when we substitute for ξ_i.

(b) **Note: This part is not in the scope of this class and will not be tested, as it requires optimization knowledge covered in EE 127. If you have taken EE 127 this problem may be of interest, if not it's probably better to skip it.**

The dual of the standard soft-margin SVM is:

$$\max_{\alpha} \alpha^\top \mathbf{1} - \frac{1}{2}\alpha^\top \mathbf{Q} \alpha \tag{12}$$

$$s.t. \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

$$0 \leq \alpha_i \leq C \quad i = 1, \cdots, n \tag{14}$$

where $\mathbf{Q} = (\operatorname{diag} \mathbf{y})\mathbf{X}\mathbf{X}^T(\operatorname{diag} \mathbf{y})$

What's the dual form of the SVM with custom margin? Show the derivation steps in detail.

Solution: We start from the constrained primal problem.

$$\min_{\mathbf{w}, \alpha, \xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \phi_i \xi_i \tag{15}$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha) \geq 1 - \xi_i \quad \forall i \tag{16}$$

$$\xi_i \geq 0 \quad \forall i \tag{17}$$

Using α and β for our dual variables, the Lagrangian is then

$$\mathcal{L}(\mathbf{w}, \alpha, \xi, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \phi_i \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha)) + \sum_{i=1}^{n} \beta_i(-\xi_i) \tag{18}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i(\mathbf{w}^\top \mathbf{x}_i - \alpha) + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \beta_i)\xi_i \tag{19}$$

The optimization we want to solve then is

$$\max_{\alpha\geq 0,\beta\geq 0}\ \min_{\mathbf{w},\alpha,\xi_i} \mathcal{L}(\mathbf{w},\mathbf{b},\xi,\alpha,\beta).$$

Since the problem is convex and strictly feasible, we know that the KKT conditions must hold for the dual optimal solutions. We will now use the KKT conditions to simplify our problem. First, we know that the gradients with respect to all primal variables must be 0 by the stationarity condition. From this, we have that

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w}^* - \sum_{i=1}^{n} \alpha_i^* y_i x_i = 0 \implies \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

$$\nabla_{\alpha}\mathcal{L} = \sum_{i=1}^{n} \alpha_i^* y_i = 0.$$

$$\nabla_{\xi_i}\mathcal{L} = C\phi_i - \alpha_i^* - \beta_i^* = 0 \quad i = 1,\dots,n.$$

Since α,β are restricted to being greater than or equal to 0, the last equality implies that $\alpha_i^* \leq C\phi_i$. Now using the equations above, we can simplify the Lagrangian to

$$\mathcal{L}(\mathbf{w},\alpha,\xi,\alpha^*,\beta^*) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\alpha_i^* - \sum_{i=1}^{n}\alpha_i^* y_i \mathbf{w}^\top \mathbf{x}_i.$$

We need to minimize the above function then to find the optimal $\mathbf{w}^*,\alpha^*,\xi^*$.

$$\begin{aligned}
\min_{\mathbf{w},\alpha,\xi} \mathcal{L}(\mathbf{w},\alpha,\xi,\alpha^*,\beta^*) &= \mathcal{L}(\mathbf{w}^*,\alpha^*,\xi^*,\alpha^*,\beta^*) && (20)\\
&= \frac{1}{2}\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n}\alpha_i^* - \sum_{i=1}^{n}\alpha_i^* y_i(\sum_{j=1}^{n}\alpha_j^* y_j x_j)^\top \mathbf{x}_i && (21)\\
&= \frac{1}{2}\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n}\alpha_i^* - \sum_{i=1}^{n}(\sum_{j=1}^{n}\alpha_j^* y_j x_j)^\top(\alpha_i^* y_i \mathbf{x}_i) && (22)\\
&= \frac{1}{2}\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n}\alpha_i^* - \|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 && (23)\\
&= \sum_{i=1}^{n}\alpha_i^* - \frac{1}{2}\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 && (24)\\
&= \mathbf{1}^\top\alpha^* - \frac{1}{2}\alpha^*\mathbf{Q}\alpha^* && (25)
\end{aligned}$$

where we let $\mathbf{Q} = (\operatorname{diag}\mathbf{y})\mathbf{X}\mathbf{X}^T(\operatorname{diag}\mathbf{y})$.

Noting the previous constraints resulting from the KKT conditions, we then get the dual problem is the following maximization.

$$\max_{\alpha} \mathbf{1}^{\top}\alpha^* - \frac{1}{2}\alpha^*\mathbf{Q}\alpha^* \tag{26}$$

$$s.t. \quad \alpha^{\top}\mathbf{y} = 0 \tag{27}$$

$$0 \leq \alpha_i \leq C\phi_i \quad i = 1, \ldots, n \tag{28}$$

(c) **From the dual formulation above, how would you kernelize the SVM with custom margins? What role does the ϕ_i play in the kernelized version?**

Solution: We can kernelize the SVM in the same way as in the normal SVM (e.g., note that in the definition of $\mathbf{Q}$, we have an $XX^{\top}$; we replace this with the kernel matrix K).

The ϕ_i serve to adjust the constraints on α_i. If ϕ_i is very large (in the primal this means we want the margin violations to be small for the data point x_i), the constraint on α_i will be very loose. Similarly, if ϕ_i is very small (in the primal this means we allow the margin violation to be large for the data point x_i), the constraint on α_i becomes much tighter.

2 SVMs for Novelty Detection

This problem is an SVM-variant that works with training data from only one class.

The classification problems we saw in class are two-class or multi-class classification problems. What would one-class classification even mean? In a one-class classification problem, we want to determine whether our new test sample is *normal* (not as in Gaussian), namely whether it is a member of the class represented by the training data or whether it is *abnormal*. One-class classification is also called *outlier detection*. In particular, we assume that all/most of the training data are from the normal class, and want to somehow model them, such that for new unseen test points, we can tell whether they "look like" these points, or whether they are different (i.e, abnormal).

For example, Netflix may want to predict whether a user likes a movie but only have thumbs-up data about movies that the user liked and no thumbs-down votes at all. How can we deal with learning with no negative training samples?

(a) Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be your training data for the one-class classification problem (all supposedly belonging to one class — the normal class). One way to formulate one-class classification using SVMs is to have the goal of finding a decision plane which goes through the origin, and for which all the training points are on one side of it. We also want to maximize the distance between the decision plane and the data points. Let the equation of the decision plane H be

$$H := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^{\top}\mathbf{x} = 0\}. \tag{29}$$

Let the margin m be the distance between the decision plane and the data points

$$m = \min_i \frac{|\mathbf{w}^{\top}\mathbf{x}_i|}{||\mathbf{w}||}. \tag{30}$$

If the convex hull of the training data does not contain the origin, then it is possible to solve the following optimization problem.

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \quad \|\mathbf{w}\|_2^2 \tag{31}$$

$$\text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i \geq 1, \quad 1 \leq i \leq n \tag{32}$$

Argue that in the above hard one-class SVM optimization (assuming that the convex hull of the training data does not contain the origin), **the resulting margin is given by** $\widehat{m} = \frac{1}{\|\widehat{\mathbf{w}}\|}$**.**

Solution: We claim that for some $j \in \{1, \ldots, n\}$, the constraint $\widehat{\mathbf{w}}^\top \mathbf{x}_j \geq 1$ holds with equality (i.e. $\widehat{\mathbf{w}}^\top \mathbf{x}_j = 1$). If this is not the case, then $\omega := \min_i \widehat{\mathbf{w}}^\top \mathbf{x}_i$ is strictly greater than 1. Thus we can make $\widehat{\mathbf{w}}$ smaller without breaking the constraints, as

$$\forall i \quad \left(\frac{\widehat{\mathbf{w}}}{\omega}\right)^\top \mathbf{x}_i = \frac{\widehat{\mathbf{w}}^\top \mathbf{x}_i}{\omega} \geq 1$$

yet

$$\left\|\frac{\widehat{\mathbf{w}}}{\omega}\right\| = \frac{\|\widehat{\mathbf{w}}\|}{\omega} < \|\widehat{\mathbf{w}}\|$$

contradicting the fact that $\widehat{\mathbf{w}}$ is optimal.

Since $\widehat{\mathbf{w}}^\top \mathbf{x}_i \geq 1$ for all i, with equality for at least one i, we have

$$\widehat{m} = \min_i \frac{|\widehat{\mathbf{w}}^\top \mathbf{x}_i|}{\|\widehat{\mathbf{w}}\|} = \frac{1}{\|\widehat{\mathbf{w}}\|}$$

as claimed.

(b) The optimal $\widehat{\mathbf{w}}$ in the hard one-class SVM optimization problem defined by (31) and (32) is identical to the optimal $\widehat{\mathbf{w}}_{\text{two-class}}$ in the traditional two-class hard-margin SVM you saw in class using the augmented training data $(\mathbf{x}_1, 1)$, $(\mathbf{x}_2, 1)$, …, $(\mathbf{x}_n, 1)$, $(-\mathbf{x}_1, -1)$, $(-\mathbf{x}_2, -1)$, …, $(-\mathbf{x}_n, -1)$.

Argue why this is true by comparing the objective functions and constraints of the two optimization problems, as well as the optimization variables.

Solution: The traditional two-class hard-margin SVM problem writes

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad 1 \leq i \leq n \end{aligned}$$

Plugging our augmented training set into the above yields

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_i + b \geq 1, \quad 1 \leq i \leq n \\ & \mathbf{w}^\top \mathbf{x}_i - b \geq 1, \quad 1 \leq i \leq n. \end{aligned}$$

Note that for any value of b, the constraints of our one-class problem ($\mathbf{w}^\top \mathbf{x}_i \geq 1$) would be satisfied, since by adding up the two constraints associated with each datapoint we obtain

$$2\mathbf{w}^\top \mathbf{x}_i = (\mathbf{w}^\top \mathbf{x}_i + b) + (\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 + 1 = 2$$

Also note that the objective function, $\frac{1}{2}\|\mathbf{w}\|_2^2$, does not contain b. Thus, we are free to choose $b = 0$, which yields the optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_i \geq 1, \quad 1 \leq i \leq n. \end{aligned}$$

which is precisely the one-class SVM problem.

(c) It turns out that the hard one-class SVM optimization cannot deal with problems in which the origin is in the convex hull of the training data. To extend the one-class SVM to such data, we use the hinge loss function

$$\max\{0, 1 - \mathbf{w}^\top \mathbf{x}_i\} \tag{33}$$

to replace the hard constraints used in the one-class SVM so that the optimization becomes

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \max(0, 1 - \mathbf{w}^\top \mathbf{x}_i). \tag{34}$$

Explain how the hyper-parameter $C > 0$ affects the behavior of the soft one-class SVM in (34)**.**

Solution: When C is small, the one class SVM will optimize toward maximizing the margin while allowing more samples in the training data to be implicitly classified as outliers.

The larger C is, the harder the one class SVM will try to minimizing the number of training points implicitly classified as outliers. In the limit $C \to \infty$ we recover the hard-margin SVM, since every point must satisfy $\mathbf{w}^\top \mathbf{x}_i \geq 1$.

Finally, C controls what extent we fit (or overfit) our data. For example, if we find that the model is overfitting the training data (e.g., this could happen if there are some genuine outliers in the training data), then we could decrease C to alleviate the overfitting problem.

(d) Your friend claims that linear models like the one-class SVM are too simple to be useful in practice. After all, for the example training data in Figure 1, it is impossible to find a sensible decision line to separate the origin and the raw training data. Suppose that we believe the right pattern for "normalcy" here is everything within an approximate annulus around the unit circle. **How could you use the one-class SVM to do the right thing for outlier detection with such data? Explain your answer.**

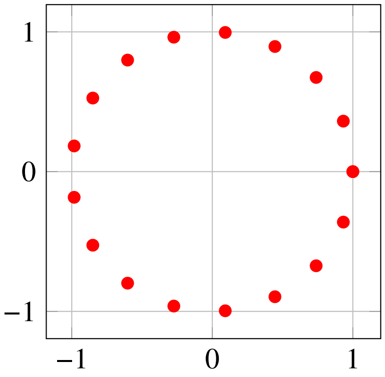

Figure 1: Counterexample provided by your friend.

Solution:

The origin is in the convex hull of the data provided by your friend, so there exists no hyperplane that can separate the origin from this data. However, we can project the data into a higher dimensional space where our training data are linearly separable from the origin. We can add explicit features to lift the problem into a space wherein an annulus can be represented as being on one side of a hyperplane.

Note that it is *not* sufficient to use a quadratic kernel, as we need polynomial features of degree at least 4. We want to be able to detect both outliers within the circle of training data and those outside the circle. A quadratic kernel will not produce a boundary that separates the annulus from both its interior and its exterior. To handle the detection we need a feature of something like $(\|x\|^2 - 1)^2$

3 Logistic posterior with exponential class conditionals

Suppose we have the job of binary classification given a scalar feature $X \in \mathbb{R}_{\geq 0}$ Now, suppose the distribution of X conditioned on the class y is exponentially distributed with parameter λ_y, i.e.,

$$\begin{aligned}
&X \in \mathbb{R}_{\geq 0} \\
&P(X = x|Y = y) = \lambda_y \exp\big(-\lambda_y x\big), \quad \text{where } y \in \{0, 1\} \\
&Y \sim \text{Bernoulli}(\pi)
\end{aligned}$$

(a) Show that the posterior distribution of the class label given X is a logistic function, however with a linear argument in X. That is, show that $P(Y = 1|X = x)$ is of the form $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax + b$ is linear in x.

(b) Assuming 0-1 loss, what is the optimal classifier and decision boundary?

Solution:

We are solving for $P(Y = 1|x)$. By Bayes Rule, we have

$$\begin{aligned}
P(Y = 1|x) &= \frac{P(x|Y = 1)P(Y = 1)}{P(x|Y = 1)P(Y = 1) + P(x|Y = 0)P(Y = 0)} \\
&= \frac{1}{1 + \frac{P(Y=0)P(x|Y=0)}{P(Y=1)P(x|Y=1)}} \\
&= \frac{1}{1 + \frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi} \exp\left(-\lambda_0 x + \lambda_1 x\right)}
\end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi} \exp\left(-\lambda_0 x + \lambda_1 x\right) = \exp\left(-(\lambda_0 - \lambda_1)x + \log\left(\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\right)\right)$$

Now we see that we have a logistic function $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax + b$ is linear (affine) in x. Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $P(Y = 1|x) > P(Y = 0|x)$. Thus, the decision boundary can be found when $P(Y = 1|x) = P(Y = 0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for x gives

$$\bar{x} = \frac{\log \frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}}{\lambda_0 - \lambda_1}.$$

If we assume $\lambda_0 > \lambda_1$, then the optimal classifier is

$$f^*(x) = \begin{cases} 1 & \text{if } x > \bar{x} \\ 0 & o.w. \end{cases}$$