

1 Support Vector Machines with Custom Margins

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

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

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

$$\xi_i \geq 0 \quad \forall i \in [1, n] \quad (3)$$

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.

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

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

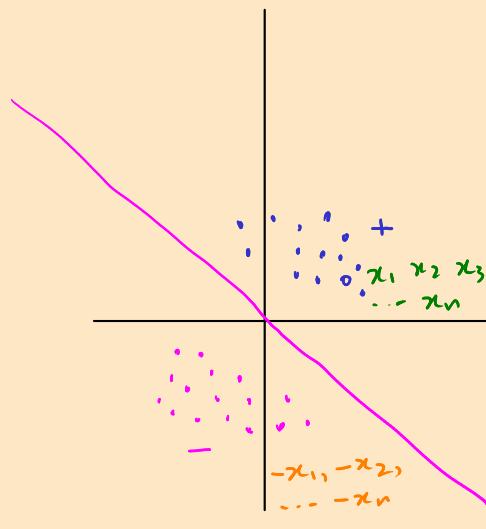
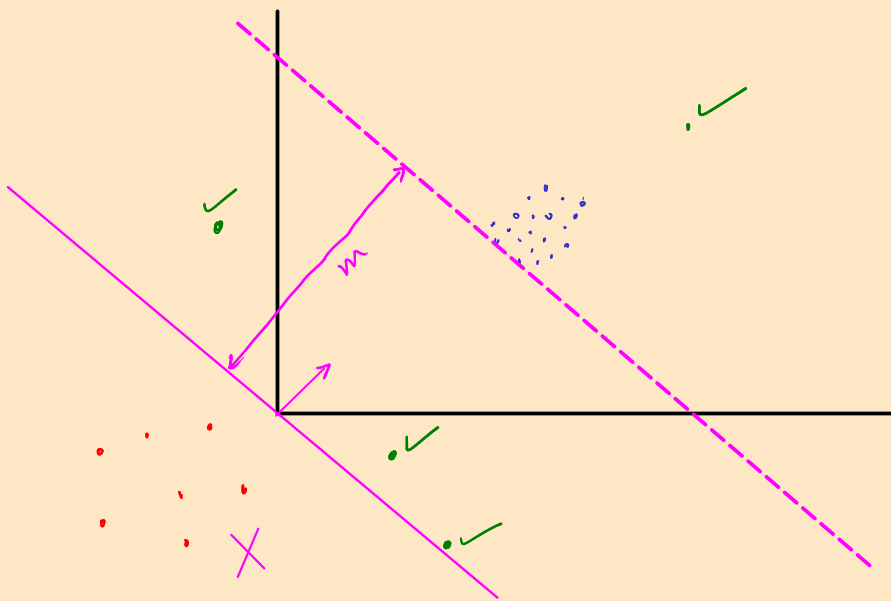
$$\xi_i \geq 0 \quad \forall i \quad (6)$$

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.

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

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



$$\min_{w, b} \frac{1}{2} \|w\|_2^2$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall 1 \leq i \leq 2n$$

$$= \min_{w, b} \frac{1}{2} \|w\|_2^2$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall 1 \leq i \leq n$$

$$y_i(w^T x_i + b) \geq 1 \quad \forall 1 \leq i \leq n$$

$$= \min_w \frac{1}{2} \|w\|_2^2$$

$$\text{s.t. } w^T x_i \geq 1 \quad \forall 1 \leq i \leq n$$

$$y_i = 1 \quad w^T x_i + b \geq 1 \quad \forall 1 \leq i \leq n$$

$$y_i = -1 \quad -1(w^T(-x_i) + b) \geq 1$$

$$\Rightarrow w^T x_i - b \geq 1 \quad \forall 1 \leq i \leq n$$

$$\square + \square \Rightarrow 2w^T x_i \geq 2 \quad \forall 1 \leq i \leq n$$

$$\Rightarrow w^T x_i \geq 1 \quad \forall 1 \leq i \leq n$$

1	2	3	...	n-1	n	n+1	n+2	...	2n
x_1	x_2	x_3	...	x_{n-1}	x_n	$-x_1$	$-x_2$...	$-x_n$

- (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:

$$\max_{\alpha} \alpha^T \mathbf{1} - \frac{1}{2} \alpha^T \mathbf{Q} \alpha \quad (8)$$

$$s.t. \sum_{i=1}^n \alpha_i y_i = 0 \quad (9)$$

$$0 \leq \alpha_i \leq C \quad i = 1, \dots, n \quad (10)$$

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

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

- (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?**

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 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 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

$$H := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^T \mathbf{x} = 0\}. \quad (11)$$

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

$$m = \min_i \frac{|\mathbf{w}^T \mathbf{x}_i|}{\|\mathbf{w}\|}. \quad \mathbf{w}^T \mathbf{x} = 0 \quad \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|} \quad (12)$$

Assume $\forall i \hat{w}^T x_i > 1$.
 $\omega = \min_i \hat{w}^T x_i > 1$

$$\|\alpha v\| = \alpha \|v\|$$

$$\hat{m} = \min_i \frac{|\hat{w}^T x_i|}{\|\hat{w}\|_2} = \frac{1}{\|\hat{w}\|_2}$$

$$\left(\frac{\hat{w}}{\omega}\right)^T x_i \geq 1$$

$$\exists i \left(\frac{\hat{w}}{\omega}\right)^T x_i = 1$$

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

$$w' = \frac{\hat{w}}{\omega}$$

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

$$\hat{w} = \arg \min_w \|w\|_2^2$$

$$\text{subject to } w^T x_i \geq 1, \quad 1 \leq i \leq n$$

$$\exists i \hat{w}^T x_i = 1$$

$$\alpha = 0$$

$$y_i(w^T x_i + \alpha) \geq 1 \quad (13)$$

$$(14)$$

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** $\hat{m} = \frac{1}{\|\hat{w}\|}$.

- (b) The optimal \hat{w} in the hard one-class SVM optimization problem defined by (13) and (14) is identical to the optimal $\hat{w}_{\text{two-class}}$ in the traditional two-class hard-margin SVM you saw in class using the augmented training data $(x_1, 1), (x_2, 1), \dots, (x_n, 1), (-x_1, -1), (-x_2, -1), \dots, (-x_n, -1)$.

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

- (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

$$\max\{0, 1 - w^T x_i\} = 0 \quad (15)$$

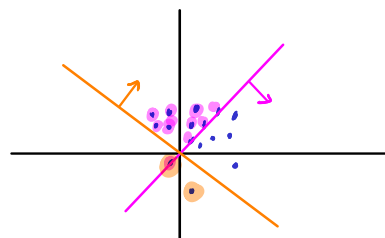
to replace the hard constraints used in the one-class SVM so that the optimization becomes

$$\hat{w} = \arg \min_w \underbrace{\frac{1}{2} \|w\|_2^2}_{\text{low hinge loss}} + C \sum_{i=1}^n \underbrace{\max(0, 1 - w^T x_i)}_{\text{high hinge loss}} \quad (16)$$

Explain how the hyper-parameter $C > 0$ **affects the behavior of the soft one-class SVM in** (16).

$$C \rightarrow 0$$

$$C \rightarrow \infty$$



- (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.**

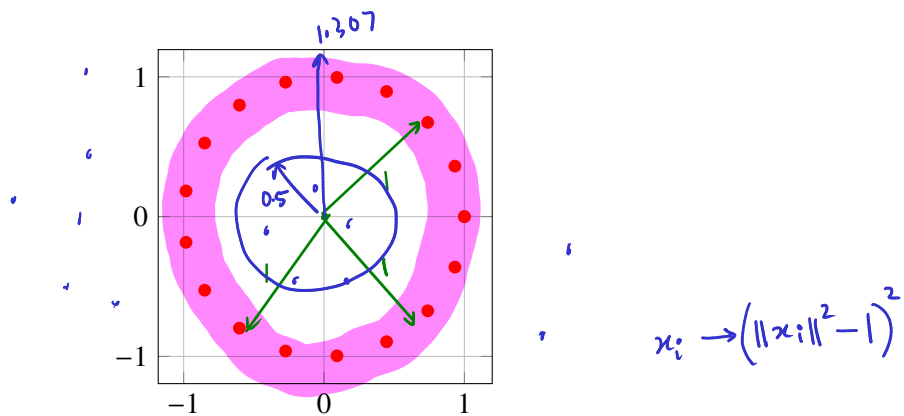
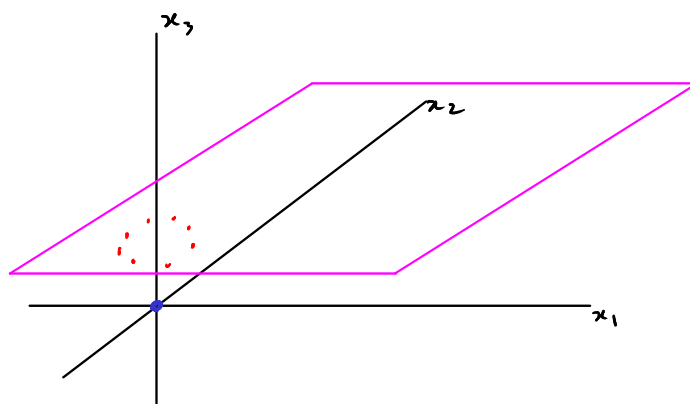


Figure 1: Counterexample provided by your friend.



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.,

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

- (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 $\frac{1}{1+\exp(-h(x))}$, where $h(x) = ax + b$ is linear in x .
- (b) Assuming 0-1 loss, what is the optimal classifier and decision boundary?