CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS3

1 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).

One practical example is malware detection. More often than not, we have a good idea of what a normal or benign program looks like (through a sequence of system calls or its binaries). However, malicious programs continuously evolve as they look for new vulnerabilities and try to evade the detector, making it particularly difficult to collect the up-to-date malware samples that we care about. Hence, as a preventive measure, malware detectors often flag suspicious programs that do not conform to the known benign ones.

(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 that 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

$$H := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} = 0\}. \tag{1}$$

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

$$m = \min_i \frac{|\mathbf{w}^\top \mathbf{x}_i|}{\|\mathbf{w}\|}. \tag{2}$$

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

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \quad \|\mathbf{w}\|_2^2 \tag{3}$$

$$\text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i \geq 1, \quad 1 \leq i \leq n \tag{4}$$

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 $j \in \{1, \dots, n\}$, 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 (3) and (4) is identical to the optimal $\widehat{\mathbf{w}}_{\text{two-class}}$ in the traditional two-class hard-margin SVM you saw in class and in the first discussion section 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

$$\max\{0, 1 - \mathbf{w}^\top \mathbf{x}_i\} \tag{5}$$

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

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \max(0, 1 - \mathbf{w}^\top \mathbf{x}_i). \tag{6}$$

Explain how the hyper-parameter $C > 0$ affects the behavior of the soft one-class SVM in (6)**.**

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.**

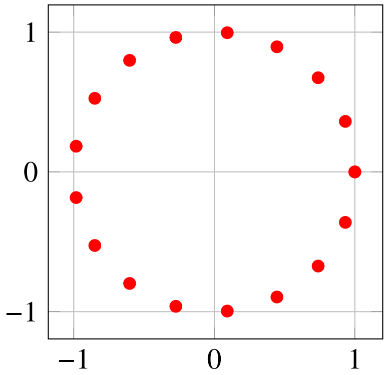

Figure 1: Counterexample provided by your friend.

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 $(\|x\|^2 - 1)^2$

2 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

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

3 Estimating Population of Grizzly Bears

An environmentalist Amy wants to estimate the number grizzly bears roaming in a forest of British Columbia, Canada. She tracks $n = 20$ bears on her first visit to the forest, and marks them with an electronic transmitter. A month later, she returns to the same forest and tracks $k = 15$ bears with only $x = 7$ having the transmitter on them. Assume that on each visit, she observes a uniformly random sample of bears.

(a) Note that the number of bears tracked during Amy's two visits n, k was chosen by her. The number of bears she found with transmitter attached is her only observation.

Assuming Amy was equally likely to encounter any of the grizzly bears during her visits, what is the likelihood $\mathcal{L}(N; x)$ of the bear population N given her observation (i.e., the number of bears with transmitter observed) x?

Solution: The likelihood $\mathcal{L}(N; x)$ is probability that Amy saw x bears with transmitters on her second trip, given N total bears. The number of ways Amy could have capture k bears on second visit $= \binom{N}{k}$. The number of ways x of them had transmitter $\binom{N-n}{k-x} \times \binom{n}{x}$. Thus, likelihood is given by:

$$\mathcal{L}(N; x) = \frac{\binom{N-n}{k-x}\binom{n}{x}}{\binom{N}{k}}$$

(b) One way to estimate the bear population is to maximize the likelihood $\mathcal{L}(N; x)$. This is called *Maximum Likelihood Estimation* (MLE), and is widely studied in statistics. Derive the expression for MLE estimate of the population $\hat{N}$ in terms of number of bears tracked in both visits (parameters n, k) and number of bears with transmitter found (observation x).

Solution: Since the random variable N is discrete, calculus isn't the best way to optimize this. Alternatively, look at the likelihood ratio $R(N|x) = \frac{\mathcal{L}(N;x)}{\mathcal{L}(N-1;x)}$. While $R(N|x) \geq 1$, likelihood increases with increasing N, and decreases if $R(N|x) \leq 1$. Thus, $R(N|x) = 1$ should be satisfied by MLE estimate $\hat{N}$.

Simplify the expression for likelihood ratio:

$$R(N|x) = \frac{\binom{N-n}{k-x}\binom{n}{x}}{\binom{N}{k}} \frac{\binom{N-1}{k}}{\binom{N-n-1}{k-x}\binom{n}{x}} = \frac{\binom{N-n}{k-x}\binom{N-1}{k}}{\binom{N-n-1}{k-x}\binom{N}{k}}$$

Simplify by using $\binom{n-1}{k}/\binom{n}{k} = \frac{n-k}{n}$ to get $R(N|x) = \frac{(N-k)(N-n)}{N(N-n-k+x)}$.

Solving $R(N|x) = 1$:

$$R(N|x) = 1 \implies (N-k)(N-n) = N(N-n-k+x)$$

$$\implies N^2 - Nk - Nn + nk = N^2 - Nk - Nn + Nx \implies nk = Nx \implies \hat{N} = \frac{nk}{x}$$

(c) What is Amy's MLE estimate $\hat{N}$ of the bear population?

Solution: $\frac{nk}{x} = \frac{15\times 20}{7} = 300/7$ is not an integer. But by the logic of likelihood ratios, the MLE estimate must be the closest integers to 300/7. The closest integers are 42 and 43. We need to evaluate the Likelihood Ratio for both of them in order to find the true MLE $\hat{N}$.

$R(42|7) = \frac{(42-15)(42-20)}{42(42-15-20+7)} = \frac{27\times 22}{42\times 14} \approx 1.01$.

$R(43|7) = \frac{(43-15)(43-20)}{43(43-15-20+7)} = \frac{28\times 23}{43\times 15} \approx 0.998$.

$R(42|7) > 1 \implies \mathcal{L}(42;7) > \mathcal{L}(41;7)$ and $R(43|7) < 1 \implies \mathcal{L}(43;7) < \mathcal{L}(42;7)$. Therefore, $\mathcal{L}(42;7)$ is the largest and $\hat{N} = 42$.

In general, the greatest integer less or equal to $\frac{nk}{x}$ is the true $\hat{N}$.

(Caution: do not attempt to calculate the actual likelihood. They involve really large numbers!)