DS603: Robust Machine Learning **Autumn 2025**

Lecture 4 — 8th August

Lecturer: Arjun Bhagoji *Scribe: Anilesh Bansal*

4.1 Unsupervised Learning for Anomaly Detection

Let $X \subseteq \mathbb{R}^d$ be the data space given by some task or application sampled from a ground-truth distribution $\mathbb{P}^+$ with a corresponding pdf $p^+(x)$. We define a set of anomalies as

$$A := \left\{x \in \mathscr{X} \;\middle|\; p^+(x) \leqslant \tau\right\}, \tau \geqslant 0. \tag{1}$$

Let P be the ground-truth data-generating distribution on data space $X \subseteq \mathbb{R}^d$ with corresponding density $p(x)$, that is, the distribution that generates the observed data. For now, we assume that this data-generating distribution exactly matches the normal data distribution, that is, $\mathbb{P} \equiv \mathbb{P}^+$ and $p \equiv p+$. This assumption is often invalid in practice, of course, as the data-generating process might be subject to noise or contamination.

4.1.1 Clustering Assumption

We assume that there exists some threshold $\tau \geqslant 0$ such that

$$X \backslash A = \left\{x \in \mathscr{X} \;\middle|\; p^+(x) > \tau\right\} \tag{2}$$

is non-empty and small (in the Lebesgue measure sense, think volume). This does not imply that the full support $supp(p^+) = \{x \in \mathscr{X} \mid p^+(x) > 0\}$ of must be bounded; only that some high-density subset of the support is bounded. A standard univariate Gaussian's support is the full real axis, for example, but approximately 95% of its probability mass is contained in the interval [-1.96, 1.96].

4.1.2 Level Sets

The density level set of $\mathbb{P}$ for some threshold $\tau \geqslant 0$ is given by $C = \{x \in X \mid p(x) > \tau\}$.

For some fixed level $\alpha \in [0,1]$, the α-density level set C_α of distribution $\mathbb{P}$ is then defined as the smallest density level set C that has a probability of at least $1-\alpha$ under $\mathbb{P}$, that is,

$$\begin{aligned} C_\alpha &= \underset{C}{\operatorname{arginf}}\{\mu(C) \mid \mathbb{P}(C) \geqslant 1-\alpha\} \\ &= \{x \in \mathbb{X} \mid p(x) > \tau_\alpha\} \end{aligned} \tag{3}$$

where $\tau_\alpha \geqslant 0$ denotes the corresponding threshold and μ is typically the Lebesgue measure.

Given a level set C_α, we can define a corresponding threshold anomaly detector $c_\alpha : \mathscr{X} \to \{\pm 1\}$ as

$$c_\alpha(x) = \begin{cases} +1, & \text{if } x \in C_\alpha \\ -1, & \text{if } x \notin C_\alpha \end{cases} \tag{4}$$

Thus if $c_\alpha(x)$ is -1, we label the point x as an anomaly. Thus our task of finding anomalies reduces to the task of finding or estimating the underlying distribution p.

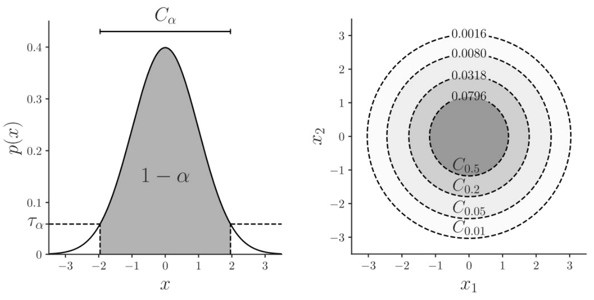

Figure 4.1. Illustration of the α-density level sets C_α with threshold τ_α for a univariate (left) and bivariate (right) standard Gaussian distribution [5]

4.2 Estimating the distribution p

4.2.1 Parametric Vs Non-Parametric density estimation

One can have priors about the distribution p and try to get estimates of parameter θ such that $p_\theta(x)$ is maximized.

For example, if it's known that samples are from a gaussian with unknown mean and known variance, $\mathcal{N}(\mu, \sigma)$, we know that $\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n}$ is an unbiased estimator of the true mean μ. In fact, it can be shown that it is the uniformly minimum-variance unbiased estimator (UMVUE) i.e. an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.

After we have an estimate of θ i.e. $\hat{\theta} = \hat{\mu}$, we can use it for estimating p and finding outliers.

However, it's not necessary that the underlying distribution can be represented through a family of distributions and parameters. In this case, it's called a non-parametric density estimation.

Thus in parametric, the θ to be estimated is finite-dimensional. In non-parametric density estimation, θ is infinite-dimensional. As a concrete example of non-parametric family of distributions, consider the family of all measurable functions, i.e. $p^+(x)$ can be any measurable density $f : \mathbb{R}^d \to \mathbb{R}$ and $\theta \equiv f$ and $\Theta \equiv \mathscr{F}$ (space of all measurable functions). In this case parameter estimation is not possible, and we do kernel density estimation instead.

4.2.2 Kernel Density Estimation

Denote by $B(x,h)$ a ball of radius h centered around x. If $p^+(x)$ changes slowly around $B(x,h)$ i.e.

$$\mathbb{P}(B(x,h)) = \int_{x \in B(x,h)} p^+(x)dx \approx p^+(x) \int_{x \in B(x,h)} dx = p^+(x)\mu(B(x,h)) \tag{5}$$

Thus we can get a local density estimate of $p^+(x)$ that depends on the choice of h as

$$\hat{p}_h^+(x) = \frac{\hat{\mathbb{P}}(B(x,h)}{\mu(B(x,h))} \tag{6}$$

where $\hat{\mathbb{P}}$ estimates the probability in a region. For example if we define $\hat{\mathbb{P}}$ as

$$\hat{\mathbb{P}}(B(x,h)) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[x_i \in B(x,h)\right] \tag{7}$$

$$\begin{aligned}
\implies \hat{p}_h^+(x) &= \frac{1}{n\mu(B(x,h))}\sum_{i=1}^{n}\mathbf{1}\left[\|x-x_i\| \leqslant h\right] \\
&= \frac{1}{nh^d V_d}\sum_{i=1}^{n}\mathbf{1}\left[\frac{\|x-x_i\|}{h} \leqslant 1\right] && (V_d = \text{volume for } d\text{-dimensional ball of unit radius}) \\
&= \frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{\|x-x_i\|}{h}\right) && (\text{Using kernel function } K(u) = \tfrac{\mathbf{1}[\|u\| \leqslant 1]}{V_d})
\end{aligned} \tag{8}$$

Here we ended up using a Box-Kernel, which is defined as above. Instead, one can also use other kernels like Gaussian ($K(u) = \frac{1}{Z}\exp(-\|u\|^2/2)$). In fact we can use any $K : \mathbb{R}^d \to \mathbb{R}$ s.t. $\int_{\mathbb{R}^d} K(u)du = 1$.

We call h as the bandwidth, and the choice of h leads to an obvious bias-variance tradeoff - if h is small $\implies$ smaller bias as a more powerful estimator is possible. However, this incurs a a larger variance since number of points used for estimation per kernel is smaller. If h is large, this variance reduces, however, then $\hat{p}_h^+(x)$ is biased.

But if we choose h appropriately, it can be shown that $\hat{p}_h^+(x)$ converges to true $p(x)$. Chen 2017 [3] considers the following three errors and gives theoretical convergence rates based on the the choice of h.

1. pointwise error i.e. $\hat{p}_h^+(x) - p(x)$
2. uniform error i.e. $\sup_x \left|\hat{p}_h^+(x) - p(x)\right|$
3. Mean Integrated Square Error (MISE) i.e. $\int \mathbb{E}\left[\hat{p}_h^+(x) - p(x)\right]^2 dx$

4.2.3 Plug-in approach to get level set estimates

We can get an estimate of the level sets as follows

$$\hat{C}_\alpha = \{x \in \mathscr{X} \mid \hat{p}_h^+(x) > \lambda\} \tag{9}$$

such that $\hat{p}_h^+(x) > \lambda$ captures some sufficient probability. However, this is a very roundabout approach of selecting outliers since we first need the density and this generates quantiles for all values of α, which is an overkill for the task of finding outliers.

Instead can we do a discriminative approach and find a function f that is +1 over some set C_α and -1 everywhere else. This is similar in ideology to just using a discriminator instead of a generative model for tasks like classification - you don't want to regenerate entire x if the final task is just classification.

4.3 Support Vector Data Description (SVDD)

Given $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathscr{X}$, consider the following constrained optimization problem:

$$\begin{aligned} \min_{R,\mathbf{c},\boldsymbol{\gamma}} \quad & R^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \gamma_i & & (10) \\ \text{subject to} \quad & \|\mathbf{x}_i - \mathbf{c}\|^2 \leqslant R^2 + \gamma_i, & i = 1, \ldots, n, \quad & (11) \\ & \gamma_i \geqslant 0, & i = 1, \ldots, n, \quad & (12) \end{aligned}$$

Where does this optimization problem arise from? We can think of $\boldsymbol{c}, R$ as the center and radius of an enclosing ball, and any test input $\boldsymbol{x}$ that lies outside this ball is deemed an outlier.

$$\|\boldsymbol{x} - \boldsymbol{c}\|^2 > R^2$$

Here ν is a hyperparameter that controls the impact of slack variables γ - intuitively it is equivalent to the fraction of points outside the enclosing ball.

We'll now look at what loss function the above problem actually minimizes and under what conditions.

4.3.1 Deriving the sphere optimization problem

Ultimately, we want a solution that minimizes the following loss function

$$\operatorname*{argmin}_h L(h) = \operatorname*{argmin}_h \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^+}\left[l(h(\boldsymbol{x}), 1)\right] + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^-}\left[l(h(\boldsymbol{x}), -1)\right] \tag{13}$$

Since we cannot get any samples from $\mathbb{P}^-$, we typically add a regularizer to the loss function to account for the latter term. Typically the regularizer is a norm constraint (say L1 or L2 norm) on the hypothesis h i.e.

$$\hat{L_R}(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(\boldsymbol{x}, 1) + R(h)$$

To get to the sphere optimization problem, we make the following assumptions -

1. Assumption 1: Define $f_\theta(\boldsymbol{x}) = R^2 - \|\boldsymbol{x} - \boldsymbol{c}\|^2$ and $h_\theta(\boldsymbol{x}) = sign(f_\theta(\boldsymbol{x}))$ where the parameter $\theta = (R, \boldsymbol{c})$. Basically we deem $\boldsymbol{x}$ as an outlier ($h_\theta = -1$) whenever $\boldsymbol{x}$ is outside the enclosing ball. And we also enforce the score f_θ drops linearly with squared norm of distance from center of the ball.

2. Assumption 2:Loss function is the shifted, cost-weighted hinge loss:

$$\ell(h_\theta(\boldsymbol{x}), y) = \begin{cases} \frac{1}{1+\nu}\max(0, -f_\theta(\boldsymbol{x})) & y = +1 \\ \frac{\nu}{1+\nu}\max(0, f_\theta(\boldsymbol{x})) & y = -1 \end{cases}$$

3. $\mathbb{P}^- = \text{Unif}(\mathscr{X})$ (this inherently assumes $\mathscr{X}$ to be bounded)

Under these assumptions, we can rewrite $L(h_\theta)$ or equivalently $L(\theta)$ as

$$L(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^+}\left[\frac{1}{1+\nu}\max\left(0, \|\boldsymbol{x} - \boldsymbol{c}\|^2 - R^2\right)\right] \to L_+(\theta)$$

$$+\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^-}\left[\frac{\nu}{1+\nu}\max\left(0, R^2 - \|\boldsymbol{x} - \boldsymbol{c}\|^2\right)\right] \to L_-(\theta)$$

Under Assumption 3:

$$
\begin{aligned}
L_{-}(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^{-}}\left[\frac{\nu}{1+\nu}\max(0, R^2 - \|\boldsymbol{x}-\boldsymbol{c}\|^2)\right] &= \frac{\nu}{1+\nu}\int \max(0, R^2 - \|\boldsymbol{x}-\boldsymbol{c}\|^2)\, d\mathbb{P}^{-}(\boldsymbol{x}) \\
&= \frac{\nu}{1+\nu}\cdot\frac{1}{\mu(\mathscr{X})}\int \max(0, R^2 - \|\boldsymbol{x}-\boldsymbol{c}\|^2)\, d\mu(\boldsymbol{x}) \\
&\leqslant \frac{\nu}{1+\nu}\cdot\frac{1}{\mu(\mathscr{X})}\int R^2\, d\mu(\boldsymbol{x}) \\
&= \frac{\nu}{1+\nu}\cdot\frac{1}{\mu(\mathscr{X})}\cdot\left(\mu(B_R(\boldsymbol{c}))\cdot R^2\right) \\
&= \frac{\nu}{1+\nu}R^2\cdot\frac{\mu(B_R(\boldsymbol{c})}{\mu(\mathscr{X})} \\
&\Rightarrow L^{-}(\theta) \leqslant \frac{\nu}{1+\nu}R^2 \\
&\Rightarrow L(\theta) \leqslant L^{+}(\theta) + \frac{\nu}{1+\nu}R^2
\end{aligned}
$$

Setting:

$$
\gamma_i \geqslant \max(0, \|\boldsymbol{x}_i - \boldsymbol{c}\|^2 - R^2) \Rightarrow L^{+}(\theta) = \frac{1}{(1+\nu)n}\sum_{i=1}^{n}\gamma_i
$$

$$
\implies L(\theta) \leqslant \frac{\nu}{1+\nu}\left[R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\gamma_i\right]
$$

Thus minimizing the upper bound on the loss, we recover the sphere optimization problem we started with

$$
L_S(\theta) = R^2 + \frac{1}{\nu n}\sum_i \gamma_i \quad \text{s.t. } \gamma_i \geqslant 0, \quad \gamma_i \geqslant \|\boldsymbol{x}_i - \boldsymbol{c}\|^2 - R^2
$$

4.4 Solving the problem - Lagrangian Dual

Let us briefly recall some definitions (see [1, Ch. 5]). For a primal problem

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\
\text{s.t.} \quad & f_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m,
\end{aligned}
$$

the *Lagrangian* is

$$
\mathscr{L}(\mathbf{x}, \boldsymbol{\alpha}) = f_0(\mathbf{x}) + \sum_{i=1}^{m}\alpha_i f_i(\mathbf{x}),
$$

with dual variables $\alpha_i \geqslant 0$. The *Lagrange dual function* is

$$
g(\boldsymbol{\alpha}) = \inf_{\mathbf{x}} \mathscr{L}(\mathbf{x}, \boldsymbol{\alpha}),
$$

and the *dual problem* is:

$$g^* = \sup_{\boldsymbol{\alpha} \geqslant 0} g(\boldsymbol{\alpha}).$$

Formally, the dual optimum should be written with sup rather than max, since in general the maximum may not be attained. In the SVDD case, however, strong duality ensures that the supremum is attained, so it is also valid to write

$$g^* = \max_{\boldsymbol{\alpha} \geqslant 0} g(\boldsymbol{\alpha}).$$

By **weak duality**, the dual optimum is always a lower bound to the primal optimum:

$$g^* \leqslant f_0(x^*).$$

4.4.1 Strong Duality

For convex optimization problems satisfying Slater's condition (strict feasibility), *strong duality* holds [1, Section 5.3]. This means the optimal primal and dual objective values coincide:

$$g^\star = f_0(x^\star)$$

Moreover, the optimal primal and dual variables correspond to a *saddle point* of the Lagrangian, i.e.,

$$\mathscr{L}(x^\star, \boldsymbol{\alpha}) \;\leqslant\; \mathscr{L}(x^\star, \boldsymbol{\alpha}^\star) \;\leqslant\; \mathscr{L}(x, \boldsymbol{\alpha}^\star), \quad \forall x,\, \forall \boldsymbol{\alpha} \geqslant 0,$$

which means that $\mathscr{L}$ is minimized in x at $x^\star$ and maximized in $\boldsymbol{\alpha}$ at $\boldsymbol{\alpha}^\star$. Equivalently, strong duality ensures the min–max equality:

$$\inf_x \sup_{\boldsymbol{\alpha} \geqslant 0} \mathscr{L}(x, \boldsymbol{\alpha}) \;=\; \sup_{\boldsymbol{\alpha} \geqslant 0} \inf_x \mathscr{L}(x, \boldsymbol{\alpha}).$$

4.4.2 Karush–Kuhn–Tucker (KKT) Conditions

For convex, differentiable objectives and constraints, the following KKT conditions must hold for optimal primal variables $\theta^\star$ and optimal dual variables $(\lambda^\star, \mu^\star)$:

1. **Primal feasibility:** All inequality and equality constraints are satisfied.
2. **Dual feasibility:** $\lambda^\star \geqslant 0$ for all inequality constraints.
3. **Complementary slackness:** For any inequality constraint $g_i(\theta) \leqslant 0$,

$$\lambda_i^\star g_i(\theta^\star) = 0.$$

4. **Stationarity:** The gradient of the Lagrangian with respect to θ vanishes:

$$\nabla_\theta \mathscr{L}(\theta^\star, \lambda^\star, \mu^\star) = 0.$$

4.4.3 Application to SVDD

The SVDD primal problem is:

$$\begin{aligned} \min_{R,\mathbf{c},\boldsymbol{\gamma}} \quad & R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\gamma_i && && (14) \\ \text{subject to} \quad & \|\mathbf{x}_i - \mathbf{c}\|^2 - R^2 - \gamma_i \leqslant 0, && i = 1,\dots,n, && (15) \\ & -\gamma_i \leqslant 0, && i = 1,\dots,n. && (16) \end{aligned}$$

HW: Verify that the SVDD primal problem is convex in (R, c, γ) and satisfies the Slater's conditions for strong duality. [2]

We form the Lagrangian by introducing multipliers $\alpha_i \geqslant 0$ for (15) and $\beta_i \geqslant 0$ for (16):

$$\begin{aligned} \mathscr{L}(R,\mathbf{c},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta}) =& R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\gamma_i \\ & + \sum_{i=1}^{n}\alpha_i\left(\|\mathbf{x}_i - \mathbf{c}\|^2 - R^2 - \gamma_i\right) - \sum_{i=1}^{n}\beta_i\gamma_i. \end{aligned} \tag{17}$$

Stationarity conditions: Taking derivatives and setting to zero:

$$\begin{aligned} \frac{\partial\mathscr{L}}{\partial R}: \quad & 2R - 2R\sum_{i=1}^{n}\alpha_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\alpha_i = 1, && (18) \\ \frac{\partial\mathscr{L}}{\partial \mathbf{c}}: \quad & -2\sum_{i=1}^{n}\alpha_i(\mathbf{x}_i - \mathbf{c}) = 0 \quad \Rightarrow \quad \mathbf{c} = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i, && (19) \\ \frac{\partial\mathscr{L}}{\partial \gamma_i}: \quad & \frac{1}{\nu n} - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i \leqslant \frac{1}{\nu n}. && (20) \end{aligned}$$

Dual problem: Substituting these into the Lagrangian yields the dual:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^\top\mathbf{x}_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbf{x}_i^\top\mathbf{x}_j && (21) \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i = 1, && (22) \\ & 0 \leqslant \alpha_i \leqslant \frac{1}{\nu n}. && (23) \end{aligned}$$

Note that in this case we can easily solve the dual problem and get the value of the center as a linear combination of the input vectors $\boldsymbol{x_i}$

$$\mathbf{c} = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i,$$

HW: Complete the proof of deriving the dual problem from the primal problem

4.4.4 Complementary Slackness for SVDD

At the optimum:

$$\alpha_i \left(\|\mathbf{x}_i - \mathbf{c}\|^2 - R^2 - \gamma_i \right) = 0, \tag{24}$$

$$\beta_i \gamma_i = 0. \tag{25}$$

These help identify the *support vectors* that lie exactly on the boundary of the enclosing ball.

4.4.5 Other Feature Spaces and the Kernel Trick

So far, everything has been formulated in the original input space $\mathbb{R}^d$, using the linear feature map

$$\phi(\mathbf{x}) = \mathbf{x}.$$

This corresponds to the kernel

$$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle.$$

Generalization: Consider a feature map

$$\phi : \mathbb{R}^d \to \mathbb{R}^p,$$

where p can be much larger than d, possibly infinite. Then, any $\mathbf{x}$ can be mapped to $\phi(\mathbf{x})$, and the SVDD formulation can be applied in this feature space.

Instead of explicitly computing $\phi(\mathbf{x})$, we can use the *kernel trick*. That is, we choose a kernel function

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \langle \tilde{\phi}(\mathbf{x}), \tilde{\phi}(\mathbf{x}') \rangle$$

for some (possibly implicit) feature map $\tilde{\phi}$, and substitute $\tilde{k}$ in place of k in the dual problem. This allows us to operate in the $\tilde{\phi}$-space without explicitly computing the mapping [6]

Examples:

1. **Polynomial kernel:**

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^d$$

2. **Gaussian kernel:**

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right)$$

Even though $\tilde{\phi}$ can be extremely high-dimensional, the kernel trick lets us compute all necessary quantities directly via $\tilde{k}$.

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Wei-Cheng Chang, Ching-Pei Lee, and Chih-Jen Lin. A revisit to support vector data description (svdd), 2013.

[3] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. 2017.

[4] Bruce E. Hansen. Lecture notes on nonparametrics. Technical report, University of Wisconsin, Spring 2009.

[5] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *arXiv preprint arXiv:2009.11732v2*, 2020. Revised version (v2) submitted Sep 28, 2020.

[6] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2002.