Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

---

1 Novelty Detection [Manzil; 60 points]

1.1 Simple Minimum Enclosing Balls

Question 1

To begin with suppose, we want to find the minimum enclosing ball for the given data of n points $x_1, ..., x_n \in \mathbb{R}^d$. That is, find the centre $z \in \mathbb{R}^d$ and the smallest radius $R \geq 0$ satisfying

$$\forall i: \quad \|x_i - z\|^2 \leq R^2 \tag{1}$$

Your goal is to state this in the form of a convex optimization problem and explain why this is a convex problem.

Let us now relax the constraint to pack all the points in the ball. How will we modify the optimization problem by introducing slack variables?
Hint: It is similar to SVM with slack variables.

Answer

The minimum enclosing ball problem is just

$$\underset{r,c}{\text{minimize}}\, r^2 \quad \text{subject to} \quad 0 \leq r,\ \|x_i - c\| \leq r. \tag{2}$$

This is literally a direct translation of "the smallest radius ball that encloses all points" into mathematics notation. It is quadratic in r and with linear constraints; hence convex. The constraint $0 \leq r$ is redundant as it is guaranteed by the other constraint.
Relaxing constraints, we can solve the problem

$$\underset{r,c}{\text{minimize}}\, r^2 + C\sum \xi_i \quad \text{subject to} \quad 0 \leq r,\ \|x_i - c\|^2 \leq r^2 + \xi_i,\ \xi_i \geq 0. \tag{3}$$

This is the same problem as before, but with some constraints allowed to be not met; the second term in the objective serves as an ℓ_1 penalty on the number of constraints not met.

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

1.2 Minimum Enclosing Balls in Feature Space

Question 2

Given n points $x_1, ..., x_n \in \mathbb{R}^d$, we want to find enclosing ball of the data in the feature space. That is, find the centre $z \in \mathbb{R}^f$ and the smallest radius $R \geq 0$ satisfying

$$\forall i: \quad \|\Phi(x_i) - z\|^2 \leq R^2 + \xi_i, \tag{4}$$

where $\xi_i \geq 0$. Also derive the dual of the optimization problem.

Answer

$$\underset{r,c,\xi}{\text{minimize}}\, r^2 + C\sum \xi_i \quad \text{subject to} \quad r \geq 0,\ r^2 + \xi_i \geq \|\Phi(x_i) - c\|^2,\ \xi_i \geq 0. \tag{5}$$

We obtain

$$L(r, c, \xi, \alpha, \eta) = r^2 + C\sum \xi_i - \sum \alpha_i(r^2 - \|\Phi(x_i) - c\|^2 + \xi_i) - \sum \eta_i \xi_i \tag{6}$$

with $\alpha, \eta \geq 0$. Derivatives in r, c, ξ need to vanish:

$$\begin{aligned} \left.\frac{\partial L}{\partial r}\right|_{r=R} &= 2R - 2\sum \alpha_i R = 0 \Rightarrow \sum_i \alpha_i = 1 \\ \left.\frac{\partial L}{\partial c}\right|_{c=z} &= \sum \alpha_i 2(\Phi(x_i) - z) = 0 \Rightarrow z = \frac{\sum \alpha_i \Phi(x_i)}{\sum \alpha_i} = \sum \alpha_i \Phi(x_i) \\ \frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \eta_i = 0 \Rightarrow \eta_i = C - \alpha_i. \end{aligned} \tag{7}$$

By the last equality, we can eliminate η and conclude that $\alpha \in [0, C]$. Notice that substituting this constraint eliminates ξ from the Lagrangian. Since $\sum \alpha_i = 1$, the r^2 terms cancel. We obtain

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_i \alpha_i \|\Phi(x_i) - \sum_j \alpha_j \Phi(x_j)\|^2 \\ &= \sum_i \left[\alpha_i \langle \Phi(x_i), \Phi(x_i) \rangle - 2\alpha_i \left\langle \Phi(x_i), \sum_j \alpha_j \Phi(x_j) \right\rangle + \alpha_i \left\langle \sum_j \alpha_j \Phi(x_j), \sum_j \alpha_j \Phi(x_j) \right\rangle \right] \\ &= \sum_i \alpha_i \langle \Phi(x_i), \Phi(x_i) \rangle - \sum_i \sum_j \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle \\ &= \sum_i \alpha_i k(0) - \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to} \quad & \alpha_i \in [0, C], \sum_i \alpha_i = 1. \end{aligned} \tag{8}$$

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

1.3 Finding Maximum Distance Separating Hyperplane

Now let us take a detour and focus on one of the method covered in the class for novelty detection, namely the method prescribed in unit 4 Linear Methods, slide 191. It is built on an idea which appears to be bizarre at first sight, i.e. finding a hyperplane in the feature, given by $f(x) = \langle w, \Phi(x) \rangle = \rho$ that has maximum distance from origin yet is still closer to the origin than the observations. Then we classify novelty for a point x, by determining which side of the hyperplane $\langle w, \Phi(x) \rangle = \rho$ does the point falls on. In this problem we will try to demystify this idea.

Question 3

Given n points $x_1, ..., x_n \in \mathbb{R}^d$ lying in a orthant, we want to find a hyperplane $\langle w, x \rangle = \rho$ that has maximum distance from origin and separating most of the points from the origin. That is, find the normal vector $w \in \mathbb{R}^d$ to the plane satisfying

$$\forall i : \langle w, x_i \rangle \geq \rho - \xi_i, \tag{9}$$

where $\xi_i \geq 0$. Your goal is to state this in the form of a convex optimization problem and explain why this is a convex problem. What if the points were mapped to feature space through $\Phi(\cdot)$, ?

Answer

To separate the data set from the origin, we solve the following quadratic program:

$$\begin{aligned} \min_{w,\rho,\xi_i} \quad & \frac{1}{2}||w||^2 - \rho + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \langle w, x_i \rangle \geq \rho - \xi_i \text{ where } \xi_i \geq 0 \; \forall i \in [n] \end{aligned} \tag{10}$$

It is convex because:

- Quadratic objective with positive definite matrix and linear terms
- Linear constraints

In case of mapping to a feature space, we have only minimal change:

$$\begin{aligned} \min_{w,\rho,\xi_i} \quad & \frac{1}{2}||w||^2 - \rho + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \langle w, \Phi(x_i) \rangle \geq \rho - \xi_i \text{ where } \xi_i \geq 0 \; \forall i \in [n] \end{aligned} \tag{11}$$

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

1.4 Establishing the Equivalence

Now we show that actually both of these problems are actually the same through some nice geometric properties of Reproducing Kernel Hilbert Space.

Question 4

Let $k(\cdot,\cdot)$ be a RBF Mercer kernel, i.e. $k(x,x') = k(x-x')$. Then show that in the feature space, any dataset of n points $x_1, ..., x_n \in \mathbb{R}^d$ will lie on surface of a hypersphere in the same orthant.

Answer

First, observe that

$$\begin{aligned} \|\Phi(x_i) - 0\|^2 &= \|\Phi(x_i)\|^2 \\ &= \langle \Phi(x_i), \Phi(x_i) \rangle \\ &= K(x_i, x_i) \\ &= k(0) \end{aligned} \tag{12}$$

which is constant; hence all points are of the same distance from the origin. This shows that all points are on the surface of a hypersphere. To show that all points in feature space are in the same orthant, it suffices to show that the angle between any two points in feature space is acute; that is, their inner product is nonnegative. To that end, note that for all i, j:

$$\begin{aligned} \langle \Phi(x_i), \Phi(x_j) \rangle &= K(x_i, x_j) \\ &= k(x_i - x_j) \\ &\geq 0. \end{aligned} \tag{13}$$

Hence all points lie in the same orthant. Note that this is not necessarily a "natural" orthant of the feature space.

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

Question 5

Derive the dual of the problem (11) and thus verify that indeed finding maximum distance separating hyperplane in the feature space is indeed equivalent to finding the minimum enclosing ball.

Answer

$$L(w,\rho,\xi;\alpha,\nu) = \frac{1}{2}\|w\|^2 - \rho + C\sum\xi_i - \sum_i \alpha_i\left[\langle w,\Phi(x_i)\rangle + \xi_i - \rho\right] - \sum_i \eta_i\xi_i \tag{14}$$

$$\begin{aligned}\frac{\partial L}{\partial w} &= w - \sum_i \alpha_i\Phi(x_i) = 0 \Rightarrow w = \sum_i \alpha_i\Phi(x_i)\\ \frac{\partial L}{\partial \rho} &= -1 + \sum_i \alpha_i = 0 \Rightarrow \sum_i \alpha_i = 1\\ \frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \eta_i = 0 \Rightarrow \eta_i = C - \alpha_i\end{aligned} \tag{15}$$

As before, $\alpha_i, \eta_i \geq 0$ implies $\alpha_i \in [0, C]$, and also as before, this substitution eliminates ξ from the Lagrangian. We then obtain the dual problem

$$\begin{aligned}\underset{\alpha}{\text{maximize}} \quad & -\frac{1}{2}\left\langle \sum_i \alpha_i\Phi(x_i), \sum_j \alpha_j\Phi(x_j)\right\rangle\\ & = -\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j K(x_i,x_j)\\ \text{subject to} \quad & \sum \alpha_i = 1, \alpha \in [0,C].\end{aligned} \tag{16}$$

Notice that this objective differs from (8) only by a constant positive factor and the term $\sum_i \alpha_i k(0)$, which is constant (equal to $k(0)$) by the second constraint. Hence, these two problems are equivalent optimization problems.

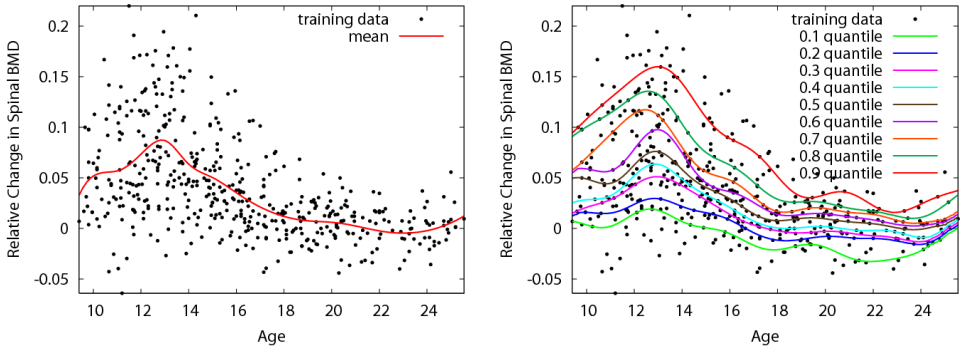
Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

---

2 Quantile Regression [Jay-Yoon; 30 points]

In this problem we consider a very similar, but a more general task to novelty detection in the setting of regression. Generally, in a typical regression setting, the prediction of y at a given x is through conditional mean. Au contraire, the desired estimate of $y|x$ is not always the conditional mean, as sometimes one may want to obtain a good estimate that satisfies the property that a proportion, τ , of $y|x$, will be below the estimate (somewhat similar to novelty detection definition). For $\tau = 0.5$ this is an estimate of the median. What might be called median regression, is subsumed under the term quantile regression (QR) where the definition of the quantile is given below.

Definition 1 *(Quantile) Denote by $y \in \mathbb{R}$ a random variable and let $\tau \in (0, 1)$. Then the τ-quantile of y, denoted by μ_τ is given by the infimum over μ for which $\Pr\{y \leq \mu\} = \tau$. Likewise, the conditional quantile $\mu_\tau(x)$ for a pair of random variables $(x, y) \in \mathcal{X} \times \mathbb{R}$ is defined as the function$\mu_\tau : \mathcal{X} \longrightarrow \mathbb{R}$ for which pointwise μ_τ is the infimum over μ for which $\Pr\{y \leq \mu|x\} = \tau$.*



(a) Conditional mean analysis (b) Conditional quantile analysis

Figure 1: Comparsion between mean and quantile regression

There exist a large area of problems where estimating quantile, rather than mean, could be helpful and following are few of the motivating examples:

- A device manufacturer may wish to know what are the 10%and 90% quantiles for some feature of the production process, so as to tailor the process to cover 80% of the devices produced.
- For risk management and regulatory reporting purposes, a bank may need to estimate a lower bound on the changes in the value of its portfolio which will hold with high probability.
- A pediatrician requires a growth chart for children given their age and perhaps even medical background, to help determine whether medical interventions are required, e.g. while monitoring the progress of a premature infant.

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

2.1 Quantile Estimator

$$l_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau-1)\xi & \text{if } \xi < 0, \end{cases} \tag{17}$$

Question 6

Prove the following properties about the minizier μ_τ of $\sum_{i=1}^m l_\tau(y_i - \mu_\tau)$:

1. The number of terms, m_-, with $y_i < \mu_\tau$ is bounded from above by τm.
2. The number of terms, m_+, with $y_i > \mu_\tau$ is bounded from above by $(1-\tau)m$.
3. For $m \to \infty$, the fraction $\frac{m_-}{m}$, converges to τ if $\Pr(y)$ does not contain discrete components.

Answer

1. Suppose on the contrary that $m_- > \tau m$. Without loss of generality suppose that the y_i are indexed in increasing order. Denote $y_{\tau m} = y_{\lceil \tau m \rceil}$ and note that $y_{\tau m} < \mu_\tau$.

$$\begin{aligned} J(\mu_\tau) &= \sum_{i=1}^{m_-}(1-\tau)(\mu_\tau - y_i) + \sum_{i=m_-+1}^{m} \tau(y_i - \mu_\tau) \\ &= (m_- - m\tau)\mu_\tau + \sum_{i=1}^{m}\tau y_i - \sum_{i=1}^{m_-} y_i \\ &> (m_- - m\tau)y_{\tau m} + \sum_{i=1}^{m}\tau y_i - \sum_{i=1}^{\tau m} y_i \\ &> (\lfloor \tau m \rfloor - \tau m)y_{\tau m} + \sum_{i=1}^{m}\tau y_i - \sum_{i=1}^{\tau m} y_i \\ &= J(y_{\tau m}). \end{aligned} \tag{18}$$

   This is a contradiction; it implies μ_τ is not the minimizer.

2. This follows from applying part 1 to the $\mu_{1-\tau}$ of $-Y$.

3. If $\Pr(y)$ is not discrete, then $\forall i \Pr(y_i = \mu\tau)$ is almost surely 0, and hence $\frac{m_- + m_+}{m} \to 1$ almost surely as $m \to \infty$. Combining this with $\frac{m_-}{m} \leq \tau$ and $\frac{m_+}{m} \leq (1-\tau)$ from above, we obtain $\tau \geq \frac{m_-}{m} \approx 1 - \frac{m_+}{m} \geq \tau$. Hence for large enough m, $\frac{m_-}{m} = \tau$ almost surely.

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

2.2 Quantile Risk Optimization

The τ-quantile regression function can be found by solving the following optimization problem

$$\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}||w||_2^2. \tag{19}$$

Question 7

Show that the minimization problem (19) is equivaelent to the following minimization problem (20):

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^{(*)}} \quad & \frac{1}{m} \sum_{i=1}^{m} \left(\tau\xi_i + (1-\tau)\xi_i^{(*)} \right) + \frac{\lambda}{2}||w||^2 \\ \text{subject to} \quad & y_i - \langle \Phi(x_i), w \rangle - b \leq \xi_i \text{ and} \\ & \langle \Phi(x_i), w \rangle + b - y_i \leq \xi_i^{(*)} \text{ where } \xi_i, \xi_i^{(*)} \geq 0 \end{aligned} \tag{20}$$

Also derive the dual of the problem (20). Use $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ for notation.

Homework 7 Solution
Thanks to John A.W.B. Costanzo for providing baseline

Answer

(Equivalence.) We will argue the following arguement which guarantees the equivalence between (19) and (20):

$$\text{for } \xi_i \in \Xi, \xi_i^{(*)} \in \Xi^{(*)}, \quad \tau\xi_i + (1-\tau)\xi_i^{(*)} = l_\tau(y_i - f(x_i)) \tag{21}$$

where Ξ_i, $\Xi_i^{(*)}$ signify the given constraints in the (20). When the following condition on ξ_i and ξ_i^* holds, it is trivial that (21) holds.

$$\xi_i = \begin{cases} y_i - f(x_i) & y_i - f(x_i) > 0 \\ 0 & otherwise \end{cases} \text{ and } \xi_i^* = \begin{cases} 0 & otherwise \\ f(x_i) - y_i & y_i - f(x_i) < 0. \end{cases} \tag{22}$$

Now, we want to connect above fact to the actual optimization function, namely let's focus on the first part of (20),

$$\min_{w,b,\xi_i \in \Xi, \xi_i^{(*)} \in \Xi^{(*)}} \quad \frac{1}{m}\sum_{i=1}^{m}\left(\tau\xi_i + (1-\tau)\xi_i^{(*)}\right) \tag{23}$$

We will show that above optimization function is equivalent to original one when ξ_i, ξ_i^* follows constraints Ξ, $\Xi^{(*)}$. First, Ξ, $\Xi^{(*)}$ can be easily satisfied by setting $-\xi_i^* < 0,\ \xi_i > 0$ as the minimum, maximum value of $y_i - f(x_i)$, respectively, i.e. $-\xi_i^* \leq y_i - f(x_i) \leq \xi_i$.
Second, to observe, $\xi_i = y_i - f(x_i)$ and $\xi_i^* = -(y_i - f(x_i))$ has to be satisfied in the optimal solution, let's examine the optimization function (23). As the objective function is affine and as we are optimizing over ξ_i, when $y_i - f(x_i) > 0$, ξ_i will become $y_i - f(x_i)$ at the optimal solution as it is the minimum value it can take. Likewise, $\xi_i^* = f(x_i) - y_i$ at the optimal solution on the feasible space.
The proof completes as we have proven that under the constraints Ξ, $\Xi^{(*)}$, $\min_{w,b,\xi_i \in \Xi, \xi_i^{(*)} \in \Xi^{(*)}} \frac{1}{m}\sum_{i=1}^{m}\left(\tau\xi_i + (1-\tau)\xi_i^{(*)}\right) = \min_{w,b,\xi_i \in \Xi, \Xi^{(*)}} \frac{1}{m}\sum_{i=1}^{m}\left(\tau\xi_i + (1-\tau)\xi_i^{(*)}\right)$.
Thus, at the optimum $\xi_i = y_i - f(x_i)$ if positive and zero otherwise, and $\xi_i^{(*)} = -(y_i - f(x_i))$ if $y_i - f(x_i)$ is negative and zero otherwise. The combination of these piecewise functions in (21) is the pinball loss function.
(Continued on next page.)

Carnegie Mellon University

Homework 7 Solution

Thanks to John A.W.B. Costanzo for providing baseline

Answer

(Dual.) The Lagrangian can be written as:

$$\begin{aligned} L(w,b,\xi,\xi^{(*)},\alpha,\beta,\eta,\gamma) = \frac{1}{m}\sum_{i=1}^{m}\Big\{\tau\xi_i + (1-\tau)\xi_i^{(*)} + \alpha_i(y_i - \langle\Phi(x_i),w\rangle - b - \xi_i) \\ +\beta_i\langle\Phi(x_i),w\rangle + b - y_i - \xi_i^{(*)}) - \eta_i\xi_i - \gamma_i\xi_i^{(*)}\Big\} + \frac{\lambda}{2}||w||^2 \end{aligned} \tag{24}$$

where $\alpha_i, \beta_i, \eta_i, \gamma_i > 0$.
The derivatives of the Lagrangian are

$$\begin{aligned} \frac{\partial L}{\partial \xi_i} &= \tau - \alpha_i - \eta_i = 0 \Rightarrow \eta_i = \tau - \alpha_i, \quad \alpha_i \in [0,\tau] \\ \frac{\partial L}{\partial \xi_i^{(*)}} &= 1 - \tau - \beta_i - \gamma_i = 0 \Rightarrow \gamma_i = 1 - \tau - \beta_i, \quad \beta_i \in [0, 1-\tau] \\ \frac{\partial L}{\partial w} &= \lambda w + \frac{1}{m}\sum(\beta_i - \alpha_i)\Phi(x_i) = 0 \Rightarrow w = \frac{1}{\lambda m}\sum(\alpha_i - \beta_i)\Phi(x_i) \\ \frac{\partial L}{\partial b}\sum(\beta_i - \alpha_i)b &= 0 \Rightarrow \sum\beta_i = \sum\alpha_i \end{aligned} \tag{25}$$

Substituting these constraints into the Langrangian yields the dual problem:

$$\underset{\alpha,\beta}{\text{maximize}} \quad \frac{\lambda}{2}\|w\|^2 - \frac{1}{m}\sum(\beta_i - \alpha_i)\left[\langle\Phi(x_i),w\rangle - y_i\right]. \tag{26}$$

As β_i, α_i always appear together, if we replace them with $\zeta_i = \frac{1}{\lambda m}(\alpha_i - \beta_i)$, the problem simplifies to

$$\begin{aligned} \underset{\zeta}{\text{maximize}} \quad & \frac{1}{2}\|w\|^2 - \sum\zeta_i\left[\langle\Phi(x_i),w\rangle - y_i\right] \\ & = \frac{1}{2}\|w\|^2 - \langle w,w\rangle + \zeta^T y \\ & = -\frac{\lambda}{2}\langle w,w\rangle + \zeta^T y \\ \iff -\underset{\zeta}{\text{minimize}} \quad & \frac{1}{2}\zeta^T K\zeta - \zeta^T y \end{aligned} \tag{27}$$

subject to $\zeta \in \left[\frac{\tau-1}{\lambda m}, \frac{\tau}{\lambda m}\right]$, $\zeta^T\mathbf{1} = 0$.