Homework 3

Joseph Camacho

3.1.1

1

First off, $\rho_\tau(z)$ is a convex function, since

$$\rho_\tau(z) = \max(z\tau, z(\tau - 1))$$

can be written as the maximum of two convex functions. Similarly,

$$f(w) = \sum_i \rho_t(y_i - w)$$

is convex, being the sum of convex functions. This means that any local minimum of f is also a global minimum. Now, note that

$$\rho'_\tau = \begin{cases} \tau - 1 & \text{if } z < 0 \\ \tau & \text{otherwise.} \end{cases}$$

Notice that at $w = y_\tau$,

$$\begin{aligned} f'(y_\tau) &= \sum_i \rho'_\tau(y_i - w) \\ &= \left(\sum_{y_i < w} \rho'_\tau(y_i - w) \right) + \left(\sum_{y_i \geq w} \rho'_\tau(y_i - w) \right) \\ &= N\tau(\tau - 1) + N(1 - \tau)\tau \\ &= 0, \end{aligned}$$

where N is the number of data points. Thus, y_τ is a local minimum of f, and hence a global minimum as well. So,

$$argmin_w \sum_i \rho_t(y_i - w) = y_\tau.$$

2

It's half the L-1 loss. It's also half the absolute value.

3

Setting

$$u_i = \max(0, y_i - x_i^T\beta) \quad \text{and} \quad v_i = \max(0, x_i^T\beta - y_i)$$

shows that the minimum of

$$f(u, v) = u^T 1\tau + v^T 1(1 - \tau)$$

subject to

$$X^T\beta - y + u - v = 0$$

is no greater than the minimum of $\sum_{i=1}^{N} \rho_\tau(y_i - x_i^T\beta)$. The other direction follows from the fact that the only way to diminish $f(u, v)$ is to decrease the value of either u_i or v_i in some coordinate i. However, any decrease in u_i must result in an identical decrease in v_i, but at least one of these is already 0, so they can't be decreased any further.

4

The problem is equivalent to finding

$$L(\beta, u, v, \lambda, \mu, \nu) = \min_{\beta,u,v} \max_{\lambda,\mu,\nu} \quad \begin{array}{c} u^T 1\tau + v^T 1(1-\tau) \\ -\lambda^T(X^T\beta - y + u - v) \\ -\mu^T u - \nu^T v \end{array}$$

subject to $\mu, \nu \geq 0$. By the minimax theorem, this is the same as

$$\max_{\lambda,\mu,\nu} \min_{\beta,u,v} \quad \begin{array}{c} u^T 1\tau + v^T 1(1-\tau) \\ -\lambda^T(X^T\beta - y + u - v) \\ -\mu^T u - \nu^T v \end{array}$$

Notice that $X\lambda$ must equal zero, or the inner minimization will equal $-\infty$ (by making β arbitrarily large). So, λ is in the null-space of X. Plugging this in yields that, equivalently, we want to find a saddle point of
Notice that $X\lambda$ must equal zero, or the inner minimization will equal $-\infty$ (by making β arbitrarily large). So, λ is in the null-space of X. Plugging this in yields that, equivalently, we want to find a saddle point of

$$L_2(u, v, \mu, \nu) = u^T 1\tau + v^T 1(1-\tau) - \lambda^T(-y + u - v) - \mu^T u - \nu^T v.$$

Taking derivatives with respect to u and v and setting them equal to 0 shows that a saddle point occurs when

$$\mu = 1\tau - \lambda \quad \text{and} \quad \nu = 1(1-\tau) + \lambda.$$

The constraints $\mu \geq 0$ and $\nu \geq 0$ require that

$$\lambda \leq 1\tau \quad \text{and} \quad \lambda \geq 1(\tau - 1).$$

Plugging this in gives the equivalent maximization of

$$L_3(\lambda) = \lambda^T y$$

subject to $\tau - 1 \leq \lambda_i \leq \tau$ for all i, and λ is in the null-space of X. Setting

$$z = \lambda + 1(1-\tau)$$

gives the formulation

$$\max_z z^T y, \quad \text{subject to} \quad Xz = (1-\tau)X1, z \in [0, 1]^N,$$

as desired.

5

Using complementary slackness,

$$\lambda_i(y_i - x_i^T\beta - u_i + v_i) = 0$$

for all i. Since $\mu_i = \tau - \lambda_i$ and $\nu_i = 1 - \tau + \lambda_i$, this can be simplified to

$$\lambda_i(y_i - x_i^T\beta + 1 - 2\tau + 2\lambda_i) = 0.$$

Replacing $\lambda_i = z_i + \tau - 1$ gives

$$(z_i + \tau - 1)(y_i - x_i^T\beta + 2z_i - 1) = 0.$$

Thus, if $z_i = 0$,

$$y_i - x_i^T\beta - 1 = 0 \Longrightarrow y_i - x_i^T\beta = 1.$$

If $z_i = 1$, then $y_i - x_i^T\beta = -1$.
If $z_i \in (0, 1)$, then there are two cases to consider:

- If $z_i = 1 - \tau$, then $y_i - x_i^T\beta$ can be anything.
- Otherwise, $y_i - x_i^T\beta = 1 - 2z_i$.

6

```
tau = 0.75 | slope = 0.21740773372818534 | intercept = 0.4492025040398797
tau = 0.5 | slope = 0.23820824709390254 | intercept = -0.3307244016150209
tau = 0.25 | slope = 0.22136016726775787 | intercept = -0.9926835520240038
Copy
```

Look at the code for more details.

3.2.1

1

By Bayes' theorem,

$$P(y^* \mid X^*, X, y) = \frac{P(y^*, y \mid X^*, X)}{P(y \mid X^*, X)}.$$

The numerator and denominator on the right are

$$\frac{\exp(-\frac{1}{2}\begin{bmatrix} y \\ y^* \end{bmatrix}\ k([X \quad X^*], [X \quad X^*])^{-1}\ [y^T \quad (y^*)^T]}{\sqrt{(2\pi)^{n+m}\ |k([X \quad X^*], [X \quad X^*])|}}$$

and

$$\frac{\exp(-\frac{1}{2}y^T\ k(X, X)^{-1}\ y)}{\sqrt{(2\pi)^n\ |k(X, X)|}},$$

respectively. Dividing them gives a new normal distribution. Using the block matrix inverse for $k([X \quad X^*], [X \quad X^*])^{-1}$ lets us calculate that the mean and covariance matrix of this new normal distribution are

$$k(X^*, X)k(X, X)^{-1}y$$

and

$$k(X^*, X^*) - k(X^*, X)k(X, X)^{-1}k(X, X^*),$$

respectively.

3.3

1

This doesn't make sense... If $f \in \mathcal{F}(\mathcal{X})$ then $2f$ and $-f$ are also in $\mathcal{F}(\mathcal{X})$. So, the supremum doesn't exist. Did you intend to define $\mathcal{F}(\mathcal{X})$ as a bounded space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$, not a space of bounded functions?

The empircal version of this statement would be

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{m} \sum_{i=1}^{m} f(y_i) \right).$$

2

Your definition of ϕ is wrong. It should be a map from $\mathcal{X}$ to $\mathcal{F}$, not to $\mathbb{R}$.

Replacing $f(x)$ with $\langle f, \phi(x) \rangle_{\mathcal{H}}$ yields

$$MMD^2[\mathcal{F}, p, q] = \left(\sup_{f \in \mathcal{F}} \langle f, \mathbb{E}[\phi(x)] - \mathbb{E}[\phi(y)] \rangle \right)^2.$$

By the Cauchy-Schwartz inequality, this is less than or equal to

$$\| f \|^2 \| \mathbb{E}[\phi(x)] - \mathbb{E}[\phi(y)] \|^2 \leq \| \mathbb{E}[\phi(x)] - \mathbb{E}[\phi(y)] \|^2.$$

3

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{m} \sum_{i=1}^{m} \phi(y_i) \right\|^2$$

Letting $k(u, v) = \langle \phi(u), \phi(v) \rangle$, expanding this gives the kernel method of estimating the MMD:

$$\frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) - 2k(x_i, y_j) + k(y_i, y_j).$$

4

The calculated empirical MMD is 0.00082. My corresponding conclusion is that x and y were drawn from the same distribution--the uniform distribution on $[0, 1]$.

3.4

1

One Lagrangian function is

$$L(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \lambda_j f_j(x).$$

2

Let

$$S = \{x \mid g(x) \le 0 \text{ and } f(x) = 0\}.$$

Then

$$\overline{L}(\lambda) = \inf_x L(x, \lambda) \le \inf_{x \in S} L(x, \lambda) \le \inf_{x \in S} f(x) = f(x^*).$$

Also,

$$\sup_{\lambda_i \ge 0} \overline{L}(\lambda) \le \sup_{\lambda_i \ge 0} f(x^*) = f(x^*).$$

3

If (x^*, λ^*) is a saddle point, then

$$L(x^*, \lambda) \le L(x^*, \lambda^*)$$

for all $\lambda \in \mathbb{R}^m_{\ge 0} \times \mathbb{R}^k$. Suppose that $f_j(x^*) \ne 0$ for some j. Then consider the limit as λ_j goes to $\pm\infty$ (with the sign chosen such that $\lambda_j f_j(x^*) \to +\infty$), fixing everything else. This would cause the LHS to go to $+\infty$, which contradicts the fact that $L(x^*, \lambda)$ is bounded above by $L(x^*, \lambda^*)$.

Likewise,

$$\lambda_i^* g_i(x^*) = 0$$

for all i, because if $g_i(x^*) > 0$, then λ_i can be chosen so that the LHS goes to $+\infty$. Thus, $g_i(x^*) \le 0$ for all i, which makes $\lambda_i^* g_i(x^*) \le 0$. Equality can be achieved by setting λ_i^* to 0, and increasing $\lambda_i g_i(x^*)$ will only increase $L(x^*, \lambda)$, so equality will occur for the optimal value. Combining this together reveals that

$$L(x^*, \lambda^*) = f(x^*).$$

4

Let x' be the optimum of the primal. Then $f(x') \leq f(x^*)$. On the other hand, combining part (3) with part (1) reveals that

$$f(x^*) = L(x^*, \lambda^*) = \inf_x L(x, \lambda^*) \leq f(x').$$

Thus, $f(x^*) = f(x')$, and so x^* is an optimum of the primal.

5

The KKT conditions are

1. Stationary: The derivative of the Lagrangian with respect ot x is 0. I.e.

$$f'(x) + \sum_i \lambda_i^* g_i'(x^*) + \sum_j \lambda_j^* f_i'(x^*) = 0.$$

2. Primal feasibility. $g_i(x^*) \leq 0$ for all i and $f_j(x^*) = 0$ for all j.
3. Dual feasibility: $\lambda_i^* \geq 0$ for all i.
4. Complementary slackness:

$$\sum_{i=1}^m \lambda_i^* g_i(x^*) = 0.$$

6

Assuming primal feasibility, dual feasibility, and complementary slackness, we have

$$L(x^*, \lambda^*) = f(x^*).$$

On the other hand,

$$L(x^*, \lambda) = f(x^*) + \sum_i \lambda_i g_i(x^*) + \sum_j \lambda_j f_j(x^*).$$

Since $f_j(x^*) = 0$, this is equal to

$$f(x^*) + \sum_i \lambda_i g_i(x^*).$$

Since $g_i(x^*) \leq 0$, this is less than or equal to $f(x^*)$, as desired.

7

Since $\lambda_i \geq 0$, $\lambda_i g_i$ is a convex function. The sum of convex functions and affine functions is again a convex function. Therefore,

$$\ell(x) \triangleq L(x, \lambda^*)$$

is a convex function. The stationary condition implies that that $\ell(x^*)$ is the minimum value of $\ell(x)$ (since the derivative of a convex function is only 0 at the minimum). This means that $L(x^*, \lambda^*) \leq L(x, \lambda^*)$ for all x.