Combined

Week 2

By James Camacho

Part 1

1.1, Quantile Regression

1. Note that
$$\frac{\mathrm{d}}{\mathrm{d}x}\rho_\tau(y_i - x) = \begin{cases} -\tau & w < y_i \\ 1-\tau & w > y_i. \end{cases}$$
Let
$$f(x) = \sum_i \rho_\tau(y_i - x),$$
which is differentiable everywhere except when $x = y_i$ for some y_i, and has derivative
$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) = \sum_i I(w > y_i) - N\tau.$$
This derivative is positive when $x > y_\tau$ and negative when $x < y_\tau$ so the minimum of $f(x)$ occurs when $x = y_\tau$.
2. It's equivalent to the one-norm or absolute value, just halved. It will find the median of the data.
3. If we set
$$u_i = \begin{cases} y_i - x_i^T\beta & x_i^T\beta \le y_i \\ 0 & x_i^T\beta > y_i, \end{cases}$$
and
$$v_i = \begin{cases} 0 & x_i^T\beta \le y_i \\ x_i^T\beta - y_i & x_i^T\beta > y_i, \end{cases}$$
then
$$\sum_{i=1}^N \rho_\tau(y_i - x_i^T\beta) = u^T 1\tau + v^T 1(1-\tau),$$
$u, v \ge 0$, and
$$X^T\beta - y + u - v = 0.$$

So

$$\arg\min_{\beta\in\mathbb{R}^K}\sum_{i=1}^{N}\rho_\tau(y_i-x_i^T\beta)\geq\arg\min_{\beta,u,v}u^T1\tau+v^T1(1-\tau)$$

given $X^T\beta-y+u-v=0;u,v\geq 0$. Also any β from the RHS can be plugged into the LHS, so the reverse inequality is true as well. The two problems are equivalent.

4. We want to minimax the Lagrangian:

$$\max_{a,b,\lambda}\min_{u,v,\beta}u^T1\tau+v^T1(1-\tau)-\lambda^T(X^T\beta-y+u-v)-a^Tu-b^Tv$$

where $a,b\geq 0$. Taking a gradient w.r.t. β gives

$$\lambda^TX^T=0.$$

Taking gradients w.r.t. u,v give

$$a+\lambda=1\tau,$$

$$b-\lambda=1-1\tau.$$

Plugging this back in gives the maximization problem

$$\max_{\lambda}\lambda^Ty.$$

subject to $\lambda^TX^T=0$. If we let $z=1-1\tau+\lambda$ we get the equivalent problem

$$\max_{z}y^Tz,\quad\text{subject to } Xz=(1-\tau)X1.$$

Note that $z=1-1\tau+\lambda=1-a=b$, so $0\leq z\leq 1$ or $z\in[0,1]^n$.

5. From complementary slackness, when $z_i=0$ we have

$$a_i=1\implies u_i=0\implies y_i>x_i^T\beta.$$

Similarly, when $z_i=1$ we find

$$b_i=1\implies v_i=0\implies y_i\leq x_i^T\beta.$$

When $z_i\in(0,1)$ we get both

$$a_i,b_i>0\implies u_i=v_i=0\implies y_i=x_i^T\beta.$$

6. See code.

Part 2

2.1, Lemma from Class

We want to find $\mu = E[y^*|y]$. We have

$$E[\mu y^T] = E[E[y^*|y]y^T] = k(X^*, X),$$

and

$$E[yy^T] = k(X, X).$$

So

$$\mu = \mu y^T (yy^T)^{-1} y = E[\mu y^T] E[(yy^T)^{-1}] E[y] = k(X^*, X) k(X, X)^{-1} y.$$

We also want to find $\Sigma = E[(y^* - \mu)(y^* - \mu)^T|y]$. I've spent several days on this and haven't got a clue (well, I could use the pdf of the posterior distribution, but that would take forever to write out). I'll just take the loss on these points and look up the solution online.

Part 3

1. The empirical analogue would be to replace each expected value with the mean:
$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \frac{1}{|X|} \sum_{x \in X} f(x) - \frac{1}{|Y|} \sum_{y \in Y} f(y).$$
2. Note that
$$\mathbb{E}_{x \sim p}[f(x)] = \mathbb{E}_{x \sim p} \langle f, \phi(x) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{x \sim p}[\phi(x)] \rangle_{\mathcal{H}},$$
and similar for y, q. So
$$\text{MMD}[\mathcal{F}, p, q] = \langle f, \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \rangle_{\mathcal{H}}.$$
Squaring we get
$$\begin{aligned} \text{MMD}^2[\mathcal{F}, p, q] &= \langle f, E \rangle_{\mathcal{H}}^2, \\ &\leq \langle E, E \rangle_{\mathcal{H}}, \end{aligned}$$
where $E = \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)]$ and the inequality follows from Cauchy-Schwarz and $\langle f, f \rangle_{\mathcal{H}} \leq 1$.
3. The empirical analogue is
$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left\langle f, \frac{1}{|X|} \sum_{x \in X} \phi(x) - \frac{1}{|Y|} \sum_{y \in Y} \phi(y) \right\rangle.$$
Let $k(x, y) = \langle \phi(x), \phi(y) \rangle$ be our kernel function. From the previous problem, we have the upper bound for MMD^2:
$$\leq \langle E, E \rangle = \frac{1}{|X|^2} \sum_{x \in X} \sum_{x' \in X} k(x, x') + \frac{1}{|Y|^2} \sum_{y \in Y} \sum_{y' \in Y} k(y, y') - \frac{2}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} k(x, y).$$

4.

Part 4

I'm assuming that equality conditions are supposed to be $h_i(x) = 0 \ \forall i \in [k]$, because otherwise the notation is quite confusing.

1. The Lagrangian is
$$\begin{aligned} L(x, \lambda) &= f(x) + [g, h]^T \lambda \\ &= f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{k} \lambda_{i+m} h_i(x). \end{aligned}$$

2. As each $g_i \leq 0$ we have
$$L(x, \lambda) \leq f(x) + \sum_{i=1}^{k} \lambda_{i+m} h_i(x)$$
For the optimal x^* in the primal problem, we have $h_i(x^*) = 0$, so
$$L(x^*, \lambda) \leq f(x^*).$$
Then
$$\overline{L}(\lambda) = \inf_x L(x, \lambda) \leq L(x^*, \lambda) \leq \inf_x f(x).$$
Also,
$$\sup_{\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0} \overline{L}(\lambda) \leq \sup_{\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0} L(x^*, \lambda) \leq f(x^*).$$

3. If
$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \quad \forall \lambda \in \mathbb{R}_+^m \times \mathbb{R}^k,$$
then $\sum_{i=1}^{m} \lambda_i^* g_i(x^*) = 0$ or else we could further increase $L(x^*, \lambda^*)$ by decreasing the offending λ_i^* (where $g_i(x^*) < 0$). If any $h_i(x^*) \neq 0$, then there is no saddle point, as we can set λ_{i+m} to $\pm\infty$, so they must all equal 0. Therefore,
$$f(x^*) = L(x^*, \lambda^*)$$

4. The right hand of the saddle point gives
$$f(x^*) = L(x^*, \lambda^*) = \overline{L}(\lambda^*),$$
but from part 4.2 above we know this is a lower bound on the primal. As it is achievable, it is the optimum solution.

5. The KKT conditions are:
   (a) Stationarity: The optimum x^* satisfies $\nabla f + \lambda^T [\nabla g, \nabla h] = 0$.

(b) Primal feasibility: We need $g_i(x^*) \leq 0$ and $h_i(x^*) = 0$.
(c) Dual feasibility: We need $\lambda_i \geq 0, i \in [m]$.
(d) Complementary slackness: We need $\lambda_i g_i(x^*) = 0, i \in [m]$.

6. From primal feasibility, we have $g_i(x^*) \leq 0$ and $h_i(x^*) = 0$. So

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

with equality only when $\lambda_i g_i(x^*) = 0, i \in [m]$.

7. We are given that g, h are all convex functions (as affine is convex too). A linear combination of convex functions is convex, so L is convex in x. A bounded convex function has exactly one minima, so from dual feasibility (i.e. bounding) there is one minimum for $L(x, \lambda^*)$, which implies the right half of the saddle point condition should be satisfied.