Homework 3 Solutions
Quantile Regression, Gaussian Processes Kernels

CMU 10-715: Machine Learning (Fall 2015)
http://www.cs.cmu.edu/~bapoczos/Classes/ML10715_2015Fall/
OUT: Oct 19, 2015
DUE: Nov 2, 2015, 10:20 AM

As usual, for any programming problems we will use the following conventions:

- N is the number of datapoints, D is the dimension of each input.
- XTrain is an $N \times D$ matrix of the input data, where row i is the features for example i.
- yTrain is an $N \times 1$ vector of the input data, where the ith component is the ith output.
- XTest is an $M \times D$ matrix of the input data, where row i is the features for example i.
- yTrain is an $M \times 1$ vector of the input data, where the ith component is the ith output.

1 Quantile Regression [Eric; 35 pts]

In this section, you will derive the dual of the quantile regression problem and implement a solver.

1.1 Quantile Regression

1. (6pts) By now you may be used to minimizing problems with respect to squared error loss. Let's instead define the following loss:

$$\rho_\tau(z) = z(\tau - I(z<0)) = \begin{cases} z(\tau-1) & \text{if } z<0 \\ z\tau & \text{if } z \geq 0 \end{cases}$$

where $\tau \in (0,1)$ is called the τth quantile, and $I(z<0)$ is the indicator function that is 1 if $z<0$ and 0 otherwise. Show that

$$\min_w \sum_i \rho_\tau(y_i - w) = y_\tau$$

where y_τ is an observation sitting at the τth top percentile of the observations (specifically, this means that y_τ is at least exactly τ percent of the observations).

2. (2pts) When $\tau = 0.5$, this loss function has a well known name in statistics. What is it?

Solution:

$$\sum_i \rho_\tau(y_i - w) = \sum_{y_i<w} \rho_\tau(y_i - w) + \sum_{y_i \geq w} \rho_\tau(y_i - w) = \sum_{y_i<w} (\tau-1)(y_i-w) + \sum_{y_i \geq w} \tau(y_i - w)$$

so

$$\min_w \sum_i \rho_\tau(y_i - w) = \min_w \frac{1}{n} \sum_{y_i<w} (\tau-1)(y_i-w) + \frac{1}{n} \sum_{y_i \geq w} \tau(y_i - w)$$

Take the derivative w.r.t. w:

$$\frac{1}{n} \sum_{y_i<w} (\tau-1) + \frac{1}{n} \sum_{y_i \geq w} \tau = p_w(\tau-1) + (1-p_w)\tau = \tau - p_w$$

where p_w is the fraction of observations satisfying $y_i < w$ and n is the total number of observations. Setting this to 0 we get

$$p_w = \tau$$

so the optimal value is w that is at the τth percentile.

$$\rho_{0.5}(z) = z(0.5 - I(z < 0)) = \begin{cases} 0.5z & \text{if } z \geq 0 \\ -0.5z & \text{if } z < 0 \end{cases}$$

This is equivalent to L_1 loss.

3. (6pts) Let $\{x_i\}_{i=1,\dots,N}$ be points in $\mathbb{R}^K$ with outputs $\{y_i\}_{i=1\dots n}$ in $\mathbb{R}$. Let $X = (x_1, \dots, x_N)$. We define the regression quantile as

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau(y_i - x_i^T \beta)$$

Prove that the solution of this problem is equivalent to the solution of the following linear program. Hint: split the problem into positive and negative parts.

$$\underset{\beta \in \mathbb{R}^K, u, v \in \mathbb{R}^N}{\operatorname{argmin}} \quad u^T 1\tau + v^T 1(1-\tau), \qquad \text{subject to } X^T\beta - y + u - v = 0, u, v \geq 0$$

Solution: Define $u_i = (y_i - x_i^T\beta)_+$ and $v_i = (x_i^T\beta - y_i)_+$. Let $u = (u_1, \dots, u_N)$ and $v = (v_1, \dots, v_N)$. Then:

$$\begin{aligned} &\underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau(y_i - x_i^T \beta) \\ =&\underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau(u_i) + \rho_\tau(-v_i) \\ =&\underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} u_i\tau - v_i(\tau - 1) \\ =&\underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}}\, u^T 1\tau + v^T 1(1-\tau) \end{aligned}$$

so the primal linear program is

$$\underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}}\, u^T 1\tau + v^T 1(1-\tau), \qquad \text{subject to } X^T\beta - y + u - v = 0, u, v \geq 0$$

It will be useful to put this into standard form. Let $w = (\beta, u, v)$, $c = (0, 1\tau, 1(1-\tau))$, $Z = (X^T, I_n, -I_n)$. Then this is equivalently

$$\underset{w \in \mathbb{R}^K}{\operatorname{argmin}}\, w^T c, \qquad \text{subject to } Zw = y, u, v \geq 0$$

4. (6pts) Show that the dual of the above linear program is

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

Solution: The dual of the LP is

$$\underset{s}{\operatorname{argmax}}\, y^T s \quad \text{subject to } Z^T s \leq c$$

Note that the constraint implies that $s \leq 1\tau$ and $-s \leq 1(1-\tau)$, so $1(\tau-1) \leq s \leq 1\tau$. Thus this is equivalent to

$$\underset{s}{\operatorname{argmax}}\, y^T s \quad \text{subject to } Xs = 0, s \in [\tau - 1, \tau]^n$$

A simple change of variables $s = a - (1-\tau)1$ resutls in

$$\underset{a}{\operatorname{argmax}}\, y^T a \quad \text{subject to } Xa = (1-\tau)X1, a \in [0,1]^n$$

5. (4pts) What does the value of z_i in the dual problem tell us about $y_i - x_i^T\beta$ in the primal? Specifically, using the KKT conditions, if $z_i = 0$ then what can you say about $y_i - x_i^T\beta$? If $z = 1$? If $z \in (0,1)$?

   Solution: If $z_i = 1$, then $y_i > x_i^T\beta$. If $z_i = 0$, then $y_i < x_i^T\beta$. Otherwise, $y_i = x_i^T\beta$. These follow from the KKT conditions of complementary slackness.

6. We have generated a synthetic dataset in quantile.mat. For this problem you will use quantile regression to get the quantile estimates for this dataset.

   You should implement quantile regression by solving the primal LP. You may use any linear programming solver to do so. For example, CVXOPT (http://cvxopt.org/) is a powerful solver for general convex problems. Alternatively, you can use the glpk function in Octave (https://www.gnu.org/software/octave/doc/interpreter/Linear-Programming.html) or linprog in Matlab (http://www.mathworks.com/help/optim/ug/linprog.html).

   You may need to reformulate your problem into a canonical form accepted by the solver. Be sure to account for a non-zero intercept term. Submit the following items in your writeup:

   - (6pts) First, plot a scatterplot of the data in XTrain,yTrain. Then, plot three quantile regression lines on top of the scatterplot at the following quantiles: $\tau = 0.25, 0.50, 0.75$.
   - (3pts) Report the β values for each value of τ.
   - (2pts) Attach your code for this problem.

2 Gaussian Processes and Hyperparameter Tuning [Eric; 25pts]

2.1 Lemma from Class

1. (5pts) First, let's verify a lemma from class. Let X, y be n examples of training data and labels and let X^*, y^* be m examples of test data and labels. Let $0_n, 0_m$ denote zero vectors of length n, m respectively, and let k be some kernel function. Suppose that

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}_{\begin{bmatrix} y \\ y^* \end{bmatrix}}\left(\begin{bmatrix} 0_n \\ 0_m \end{bmatrix}, \begin{bmatrix} k(X,X) & k(X,X^*) \\ k(X^*,X) & k(X^*,X^*) \end{bmatrix}\right)$$

   Show that the posterior distribution is

$$P(y^*|X^*, X, y) = \mathcal{N}_{y^*}(\mu, \Sigma)$$

   where $\mu = k(X^*,X)k(X,X)^{-1}y$ and $\Sigma = k(X^*,X^*) - k(X^*,X)k(X,X)^{-1}k(X,X^*)$. Note: For this question, you may assume that the conditional distribution is of a Normal form, however you must derive the mean and variance.

   Solution: Let $z = y^* + Ay$ where $A = -K(X^*,X)K(X,X)^{-1}$. Note that z, y are independent since

$$\text{Cov}(z,y) = \text{Cov}(y^*,y) + \text{Cov}(Ay,y) = k(X^*,X) + Ak(X,X) = 0$$

   Then, the mean is:

$$E(y^*|X^*,X,y) = E(z - Ay|X^*,X,y) = E(z|X^*,X,y) - E(Ay|X^*,X,y) = -Ay = -K(X^*,X)K(X,X)^{-1}y$$

And the variance is:

$$\begin{aligned}
&\mathrm{Var}(y^*|X^*, X, y) \\
=&V(z - Ay|X^*, X, y) \\
=&\mathrm{Var}(z|X^*, X, y) + \mathrm{Var}(Ay|X^*, X, y) - \mathrm{Cov}(z, -Ay|X^*, X, y) - \mathrm{Cov}(-Ay, z|X^*, X, y) \\
=&\mathrm{Var}(z|X^*, X) \\
=&\mathrm{Var}(y^* + Ay|X^*, X) \\
=&\mathrm{Var}(y^*|X^*, X) + \mathrm{Var}(Ay|X^*, X) - \mathrm{Cov}(y^*, Ay|X^*, X) - \mathrm{Cov}(Ay, y^*|X^*, X) \\
=&K(X^*, X^*) + AK(X, X)A^T - \mathrm{Cov}(y^*, y|X^*, X)A^T - A\,\mathrm{Cov}(y, y^*|X^*, X) \\
=&K(X^*, X^*) + AK(X, X)A^T - K(X^*, X)A^T - AK(X, X^*) \\
=&K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)
\end{aligned}$$

2.2 GP Regression

For this problem, you will implement a basic Gaussian Process Regression. We will be using the standard radial basis kernel:

$$K(x_i, x_j) = \sigma \exp\left(\frac{-||x_i - x_j||_2^2}{2h^2}\right)$$

where σ, h are known as the scale and bandwith parameters.

For additional help, better performance, and numerical stability, we refer you to chapter 2 of Rasmussen and Williams (http://www.gaussianprocess.org/gpml/chapters/RW2.pdf).

We will test your implementation on the Concrete Compressive Strength dataset from the UCI repository. The strength of concrete is predicted from 8 features consisting of the ingredients that make up the concrete composition and its age. We have given you this dataset as an octave mat file.

We will use the following conventions for this problem:

- X1, X2 are $n_1 \times D$ and $n_2 \times D$ matrices of the input data. Note that n_1 is not necesarily equal to n_2. Each row consists of the features of a particular example.
- K is a $n_1 \times n_2$ kernel matrix for X1,X2.
- GPMean is a $N \times 1$ vector containing the predicted mean values of the GP at XTest.
- GPVariance is a $N \times N$ matrix containing the predicted covariance matrix of the GP at XTest.
- logml is a scalar value containing the log marginal likelihood of the data given the parameters.
- sigma is a the scale parameter described above, and sigmas is a $P_1 \times 1$ vector of potential parameters.
- h is a the bandwith parameter described above, and hs is a $P_2 \times 1$ vector of potential parameters.
- gamma is the noise parameter for the Gaussian Process. Specifically,

  $$\mathrm{cov}(y) = K(X, X) + \gamma I$$

1. (3pts) Implement [K] = RBFKernel(X1, X2, sigma, h), which takes as input two matrices of examples with hyperparameters sigma, h, and outputs the kernel matrix where $K_{i,j} = k(\mathsf{X1}_i, \mathsf{X2}_i)$, where k is the RBF function described above. Bonus: do this without any for loops.
2. (7pts) Implement [GPMean, GPVariance] = GPRegression(XTrain, yTrain, XTest, gamma, sigma, h), which carries out the Gaussian Process regression and returns the estimated mean and variances for the variables in XTest. See page 19 of chapter 2 in Rasmussen and Williams for help on making this computationally efficient and numerically stable.

3. (3pts) Now, we need to find hyperparameters for the Gaussian Process. One reasonable method for Gaussian processes is to choose parameters that minimizes the log marginal likelihood. First implement [logml] = LogMarginalLikelihood (XTrain, yTrain, gamma, sigma,h) which computes the log marginal likelihood of the training data given the parameters.

4. (3pts) Implement [gamma, h,sigma] = HyperParameters (XTrain, yTrain, hs, sigmas), which does a grid search across the parameters in hs,sigmas and returns the combination that minimizes the log marginal likelihood. Also set gamma to be $0.01 \cdot \sigma_y$ where σ_y is the standard deviation of the training example outputs.

5. (4pts) Run your Gaussian process regression method on the dataset provided in concrete.mat. Compare and report your results with a naive mean prediction. Get your hyperparameters by using your implemented HyperParameters functions and searching over the space of hs = logspace(-1,1,10)' * norm(std(XTrain)) and sigmas = logspace(-1,1,10)' * std(yTrain).

3 Kernel two sample-test [Fish; 40 pts]

Suppose you are collecting data on the expression level of gene No. 10715 after inserting a secret drug into mice liver. There are two labs, Lab A and Lab B, that run the experiments for you and send you the results. Of course you would hope that the environment and quality of each lab would not cause a difference in the data between the two locations. To make it simpler, assume the data from Lab A is i.i.d drawn from a disribution p, and the data from lab B are i.i.d. drawn from a distribution q. The question you would like to answer is: given data $X = \{x_1, x_2, \cdots, x_m\}$ collected from lab A and $Y = \{y_1, y_2, \cdots, y_m\}$ collected from lab B, is $p = q$?

1. (10pts) Let $\mathcal{X}$ be a sample space, and consider two distributions p and q. $p = q$ if and only if $\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{y\sim q}[f(y)]$ for all $f \in \mathcal{F}(\mathcal{X})$ where $\mathcal{F}(\mathcal{X})$ is the space of bounded continuous functions from $\mathcal{X} \to \mathbb{R}$. Using this theorem, we define the maximum mean discrepency as

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}[f(y)]\right). \quad (1)$$

To answer the question of whether $p = q$, if $\mathrm{MMD}[\mathcal{F}, p, q] = 0$, then we have $p = q$. Write the empirical version of this MMD statement that we can estimate with a dataset X, Y from the two distributions and all the functions in some $\mathcal{F}$.

Solution:

$$\hat{\mathrm{MMD}}[\mathcal{F}, p, q] := \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{m}\sum_{i=1}^{m} f(y_i)\right). \quad (2)$$

2. (10pts) The issue with the estimate from question 2 is that we need to find a sufficiently large function class to identify p and q, which is not practical. One way to solve this problem is to kernelize the function to implicitly project the data into a potentially infinite space. More importantly, using a kernel allows us to use the special properties for functions in a Reproducing Kernel Hilbert Space (RKHS): $\mathcal{H}$ is a $RKHS$ if there exists a feature mapping ϕ from space $\mathcal{X}$ to $\mathbb{R}$ such that, for all $x \in \mathcal{X}$,

$$f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}} \quad (3)$$

for every $f \in \mathcal{H}$. The subscript for the inner product indicates that the inner product is done in the RKHS instead of our sample space. Note that here f refers to the function as an object (you can imagine it as an vector in the RKHS), and $f(x) \in \mathbb{R}^d \to \mathbb{R}$ is defined over $\mathcal{X}$.

Replace $f(x)$ in (**??**) with the inner product in (**??**), and set $\mathcal{F}$ to be a unit ball in a RKHS:

$$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1, \text{ where } \|f\|_{\mathcal{H}} := \sqrt{\langle f, f\rangle_{\mathcal{H}}}\}$$

Derive an upper bound for $\mathrm{MMD}^2\left[\mathcal{F}, p, q\right]$ using $\mathbb{E}_{x\sim p}[\phi(x)]$ and $\mathbb{E}_{y\sim q}[\phi(y)]$.

Solution:

$$
\begin{aligned}
\mathrm{MMD}^2\left[\mathcal{F}, p, q\right] &= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1}\left(\mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}\left[f(y)\right]\right)\right]^2 && (4)\\
&= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1}\left(\mathbb{E}_{x\sim p}[\langle f, \phi(x)\rangle_{\mathcal{H}}] - \mathbb{E}_{y\sim q}\left[\langle f, \phi(y)\rangle_{\mathcal{H}}\right]\right)\right]^2 && (5)\\
&= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1}\langle f, \mathbb{E}_{x\sim p}[\phi(x)] - \mathbb{E}_{y\sim q}[\phi(y)]\rangle_{\mathcal{H}}\right]^2 && (6)\\
&\leq \|\mathbb{E}_{x\sim p}[\phi(x)] - \mathbb{E}_{y\sim q}[\phi(y)]\|_{\mathcal{H}}^2 && (7)
\end{aligned}
$$

3. (10pts) Replace $\mathbb{E}_{x\sim p}[\phi(x)]$ and $\mathbb{E}_{y\sim q}[\phi(y)]$ with its empirical estimates to get the kernel method of estimating MMD.

   Solution:

$$
\begin{aligned}
\hat{\mathrm{MMD}}^2\left[\mathcal{F}, p, q\right] &\leq \left\|\frac{1}{m}\sum_{i=1}^{m}\phi(x) - \frac{1}{m}\sum_{i=1}^{m}\phi(y)\right\|_{\mathcal{H}}^2 && (8)\\
&= \frac{1}{m^2}\sum_{i=1}^{m}k(x_i, x_i) - \frac{2}{m^2}\sum_{i,i'=1}^{m}k(x_i, y_{i'}) + \frac{1}{m^2}k(y_i, y_i) && (9)
\end{aligned}
$$

4. (10pts) We have provided a dataset containing two vectors drawn from some mystery distributions p and q in twosample.mat. Use the the RBF kernel to test whether the two vectors of variables have the same distribution. You can use your RBFKernel function that you wrote in question 2.1.1 with parameters $h = 10, 1, 0.1$ and $\sigma = 1$ to calculate the MMD. Use the following threshold: if MMD is less than 0.01, we say they are the same distribution, otherwise they are different.

   Report in your writeup the calculated empirical MMD and your corresponding conclusion.

4 Saddle Points in optimization[Fish; 14 pts] (Bonus)

Often we solve constrained optimization problem by first transforming it into a non-constrained optimization problem. The most common way to conduct such transformation is to introduce Lagrange multipliers and construct a dual problem for the primal problem. Before solving the dual problem, one question we would like to answer is: *Is the optimal value for the dual problem equal to the primal problem?*
Consider the convex optimization problem:

$$
\begin{aligned}
&\min f(x) && (10)\\
&\text{s.t. } g_i(x) \leq 0, \quad \forall i \in [m], && (11)\\
&\quad f_i(x) = 0 \quad \forall i \in [k], && (12)
\end{aligned}
$$

where $f_1, f_2, \cdots, f_k$ are affine functions and $f, g_1, \cdots, g_m$ are convex functions. In this question, we are going to prove that for $x^* \in \mathbb{R}$, if there exists Lagrange Multipliers $\lambda_i^* \geq 0$ such that (x^*, λ^*) is a saddle point of Lagrange function $L(x, \lambda)$, then x^* is the optimal solution for the primal problem. A point (x^*, λ^*) is said to be a saddle point of function $L(x, \lambda)$ if

$$
L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \quad \forall x \in \mathbb{R}^d, \lambda \in R_+^m \times \mathrm{R}^k. \qquad (13)
$$

1. (2pts) Introduce Lagrange Multipliers, $\lambda_1, \lambda_2, \cdots, \lambda_{k+m}$, and write out the Lagrange function for the primal problem.

   Solution:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{k} \lambda_{i+m} f_i(x), \tag{14}$$

   where $\lambda_i \geq 0$ for $i = [m]$.

2. (2pts) Show that the infimum

$$\tilde{L}(\lambda) = \inf_x L(x, \lambda) \tag{15}$$

   of the Lagrange function in $x \in X$ is a lower bound for the optimal primal value $f(x^*)$. Also prove that

$$\sup_{\lambda_1, \lambda_2, \lambda_m \geq 0} L(\lambda) \tag{16}$$

   is also a lower bound for the optimal primal value $f(x^*)$.

   Solution: Since x^* is the optimal solution, it satisfies all the constrains, i.e., $g_i(x^*) \leq 0$ and $f_i(x^*) = 0$. We say $\lambda_i \geq 0$ for $i = [m]$, so $\lambda_i g_i(x^*) \leq 0$ for all $i \in [m]$. Thus, we have

$$L(x^*, \lambda) = f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{i=1}^{k} \lambda_{m+i} f_i(x^*) \leq f(x^*). \tag{17}$$

   Since $\tilde{L}(\lambda) = \inf_x L(x, \lambda) \leq L(x', \lambda)$ for all x', we have $\tilde{L}(\lambda) \leq f(x^*)$. Since we have shown that the inequality holds for all λ that satisfies $\lambda_i \geq 0$ for $i \in [m]$, the supremum will also be the lower bound for $f(x^*)$.

3. (2pts) If (x^*, λ^*) is a saddle point of the function $L(x, \lambda)$. Prove that the left half of the saddle point conditions implies $f_i(x^*) = 0$ for $i \in [k]$ and $\sum\limits_{i=1}^{m} \lambda_i^* g_i(x^*) = 0$, so we can conclude that $f(x^*) = L(x^*, \lambda^*)$.

   Solution: The left hand part of the saddle point conditions says $\sup_{\lambda_i \leq 0 | i \in [m]} L(x^*, \lambda) \leq L(x^*, \lambda^*)$. If there is some $f_i(x^*) \neq 0$, then we can set $\lambda_i = \text{sign}(f_i(x^*))\dot{\infty}$, then the lower bound becomes $+\infty$, which contradicts the condition. So $f_i(x^*) = 0$ for $i \in [k]$. For $g_i(x^*)$, if there exists some $g_i(x) > 0$, then the lower bound goes to $+\infty$ by setting $\lambda_i = +\infty$. So $g_i(x^*) \leq 0$ for $i = [m]$. We know that $\sum\limits_{i=1}^{m} \lambda_i g_i(x) \leq 0$, and there exists a trivial solution $\lambda_i = 0$ for $i \in [m]$ such that $\sum\limits_{i=1}^{m} \lambda_i g_i(x) = 0$. So we conclude that $f(x^*) = \sup_{\lambda_i \leq 0 | i \in [m]} L(x^*, \lambda) = L(x^*, \lambda^*)$.

4. (2pts) Complete the proof by saying the right half of the saddle point condition is upper bounded by $f(x^*)$.

   Solution: Define X be the solution space that contains all the feasible points for our primal problem.

$$L(x^*, \lambda^*) \leq \inf_x L(x, \lambda^*) \leq \inf_{x \in X} L(x, \lambda^*) \leq \inf_{x \in X} f(x). \tag{18}$$

   In the previous question, we have shown that x^* in the saddle point (x^*, λ^*) satifies the constraints in the primal problem, and $L(x^*, \lambda^*) = f(x^*)$. So x^* is the solution for $\inf_{x \in X} f(x)$ since it satifies all the constraints and achieve the lower bound for $f(x)$.

5. (2pts) The other direction of the saddle point theory says that if x^* is a solution for the primal problem and the primal problem satisfies Slater C.Q., then there is a $\lambda^* \in R_+^m \times \mathrm{R}^k$ such that (x^*, λ^*) is a saddle point of $L(x, \lambda)$. We say if a problem statisfies Slater C.Q., then there is a λ^* such that (x^*, λ^*) satisfies KKT conditions. Write out the KKT conditions for the optimization problem.

Solution: Stationary:

$$-\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i \nabla g(x^*) + \sum_{i=1}^{k} \lambda_{m+i} \nabla g(x^*) \tag{19}$$

Primal feasiblility:

$$g_i(x^*) \leq 0 \quad \forall i \in [m] f_i(x^*) = 0 \quad \forall i \in [k] \tag{20}$$

Dual feasiblility:

$$\lambda_i \geq 0 \ \ \forall i \in [m] \tag{21}$$

Completmentary Slackness:

$$\lambda_i g_i(x) = 0 \quad \forall i \in [m] \tag{22}$$

6. (2pts) Use Primal feasibility, dual feasibility and complementary slackness to show the left half of the saddle point conditions.

   Solution: For all *lambda* that satisfies $\lambda_i \geq 0$ for $i \in [m]$,

$$\begin{aligned} L(x^*, \lambda) &= f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{i=1}^{k} \lambda_{m+i} f_i(x^*) \leq f(x^*) && (23) \\ &\leq f(x^*) && (24) \\ &= f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{i=1}^{k} \lambda_{m+i}^* f_i(x^*) \leq f(x^*) && (25) \\ &= L(x^*, \lambda^*), && (26) \end{aligned}$$

7. (2pts) Use dual feasibility to show the right half of the saddle point condition is a convex function in x, so the stationary condition in KKT implies that the right half of the saddle point condition should be satisfied. (Hint: Use the convexity properties we had proved in HW1.)

   Solution: Since $f, g_1, g_2, \cdots, g_m$ are convex function and $\lambda_i \geq$ for $i \in [m]$, the summation of these functions is still a convex function. Besides, $f_1, \cdots, f_k$ are affine functions, so adding them does not affect the convexity. Since right hand side is a convex funciton, the extreme value happens at a point that has gradient equals to zero. Stationary condition in KKT says that the optimal solution satisfies such property, so they are the lower bound for $L(x, \lambda^*)$.