ELE 535: Machine Learning and Pattern Recognition
Homework 4

Zachary Hervieux-Moore

Monday 15^{th} October, 2018

Exercise 1: Determine general sufficient conditions (if any exist) under which the indicated function f is convex.

a) $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = |x|$.

b) $f : (0, \infty) \to \mathbb{R}$ with $f(x) = x \ln(x)$.

c) $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) = (x^T Q x)^3$. Here $Q \in \mathbb{R}^{n\times n}$ is symmetric PSD.

d) $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) = 1 + e^{\sum_{i=1}^n |x_i|^3}$.

e) For $x \in \mathcal{C} = \{x \in \mathbb{R}^n : x_i > 0, i = 1, \dots, n\}$ let $\ln(x) = [\ln(x_i)] \in \mathbb{R}^n$ and define $f(x) = x^T \ln(x)$

Answer:

a) Simply applying the triangle inequality we get

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y| \\ \leq \lambda|x| + (1-\lambda)|y| &= \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

Thus, it is convex.

b) Here, we take the second derivative since it is continuous on the domain,

$$f''(x) = \frac{1}{x} > 0 \quad \forall x \in (0, \infty)$$

As this is always positive, then we have that f is strictly convex.

c) Here, we use theorem 7.3.2e) that states that $h(x) = g(f(x))$ is convex if g is convex and non decreasing on the range of f. Here we have $f(x) = x^T Q x$ whose range is $[0, \infty)$ as Q is PSD and $g(x) = x^3$. Thus, on $[0, \infty)$, we have that $g'(x) = 2x^2 \geq 0$ and $g''(x) = 6x \geq 0$ which means that g is on decreasing and convex respectively. Thus, we have shown that the original function is convex.

d) Again, we use the same theorem as in the previous part. First $g(x) = 1 + e^x$ and $f(x) = (\sum_{i=1}^n |x_i|)^3$. The range of $f(x)$ is $[0, \infty)$. We also have $g'(x) = e^x$ and $g''(x) = e^x$ which are both always positive and so non decreasing and convex. This shows that the original function is convex.

e) We have that $f(x) = x^T \ln(x)$ which is just a short form for

$$[f(x)]_i = x_i \ln(x_i)$$

Thus, we have

$$[\nabla f(x)]_i = 1 + \ln(x_i)$$
$$[Hf(x)]_{ii} = \frac{1}{x_i} \text{ and } [Hf(x)]_{ij} = 0 \text{ for } i \neq j$$

Since the Hessian is diagonal and always positive, then it is positive definite and hence strictly convex.

Exercise 2: You want to learn an unknown function $f : [0,1] \to \mathbb{R}$ using a set of noisy measurements (x_j, y_j), with $y_j = f(x_j) + \epsilon_j$, $j = 1, \ldots, m$. Your plan is to approximate $f(\cdot)$ by a Fourier series on $[0,1]$ with $q \in \mathbb{N}$ terms:

$$f_q(x) = \frac{a_0}{2} + \sum_{k=1}^{q} a_k \cos(2\pi k x) + b_k \sin(2\pi k x).$$

To control the smoothness of $f_q(\cdot)$, you also decide to penalize the size of the coefficients a_k, b_k more heavily as k increases.

a) Formulate the above problem as a regularized regression problem.

b) For $q = 2$, display the regression matrix, the label vector y, and the regularization term.

c) Comment briefly on how to select q.

Answer:

a) We can formulate as follows:

$$\min_{C \in \mathbb{R}^{2q+1}} \|y - X^T c\|_2^2 + \lambda \|Dc\|_2^2$$

where D is diagonal with increasing values along the diagonal and

$$c^T = \begin{bmatrix} a_0 & a_1 & b_1 & \cdots & a_q & b_q \end{bmatrix}$$
$$X^T = \begin{bmatrix} 1/2 & \cos(2\pi x_1) & \sin(2\pi x_1) & \cdots & \cos(2q\pi x_1) & \sin(2q\pi x_1) \\ \vdots & & & & & \vdots \\ 1/2 & \cos(2\pi x_m) & \sin(2\pi x_m) & \cdots & \cos(2q\pi x_m) & \sin(2q\pi x_m) \end{bmatrix}$$
$$y^T = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}$$

b) For $q = 2$, it looks like

$$\min_{a_0,a_1,a_2,b_1,b_2\in\mathbb{R}} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} 1/2 & \cos(2\pi x_1) & \sin(2\pi x_1) & \cos(4\pi x_1) & \sin(4\pi x_1) \\ \vdots & & & & \vdots \\ 1/2 & \cos(2\pi x_m) & \sin(2\pi x_m) & \cos(4\pi x_m) & \sin(4\pi x_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} \right\|_2^2$$

$$+\lambda \left\| \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} \right\|_2^2$$

c) To select q, split the data into test and training sets. Then try increasingly large q values until overfitting occurs and testing accuracy decreases.

Exercise 3: Let $D \in \mathbb{R}^{n \times n}$ be diagonal with nonnegative diagonal entries and consider the problem:

$$\min_{x \in \mathbb{R}^n} \|x - y\|_2^2 + \lambda \|Dx\|_2^2$$

This problem seeks to best approximate $y \in \mathbb{R}^n$ with a nonuniform penalty for large entries in x.

a) Solve this problem using the solution of ridge regression.

b) Show that the objective function is separable into a sum of decoupled terms. Show that this decomposes the problem into n independent scalar problems.

c) Finc the solution of each scalar problem.

d) By putting these scalar solutions together, find and interpret the solution to the original problem.

Answer:

a) We make the substitution $z = Dx$ which gives

$$\min_{z \in \mathbb{R}^n} \|D^{-1} z - y\|_2^2 + \lambda \|z\|_2^2$$

Where we assume D is invertible without loss of generality. If there are 0's on the diagonal, simply invert the parts that are non zero. Now, using our solution to ridge regression, we get

$$w_{rr}^*(\lambda) = ({D^{-1}}^2 + \lambda I_n)^{-1} D^{-1} y$$

However, since D is diagonal, then so is D^{-1} which yields the simplification

$$\begin{aligned} {w_{rr}}_i^*(\lambda) &= (\frac{1 + \lambda D_i^2}{D_i^2})^{-1} \frac{1}{D_{i\,i}} \\ &= \frac{D_i}{1 + \lambda D_i^2} y_i \end{aligned}$$

But, $x = D^{-1} z$, so $x = \frac{1}{1 + \lambda D_i^2} y_i$

b) We have

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} \|x-y\|_2^2 + \lambda\|Dx\|_2^2 \\
&= \min_{x\in\mathbb{R}^n} \sum_{i=1}^n (x_i - y_i)^2 + \lambda \sum_{i=1}^n D_i^2 x_i^2 \\
&= \min_{x\in\mathbb{R}^n} \sum_{i=1}^n (x_i - y_i)^2 + \lambda D_i^2 x_i^2
\end{aligned}$$

Which are n decoupled sums that become independent when the gradient is taken as seen in the next part.

c) Taking the gradient of the summation above and setting to 0 yields

$$\begin{aligned}
[\nabla f(x)]_i &= 2(x_i - y_i) + 2\lambda D_i^2 x_i = 0 \\
\implies x_i &= \frac{y_i}{1+\lambda D_i^2}
\end{aligned}$$

which is precisely what we found before

d) This can be written in the matrix form $x = (1 + \lambda D^2)^{-1} y$. We have that the solution of x is precisely y when either D or $\lambda = 0$ (no regularization). However, by penalizing the size of x_i, we simply scale the optimal solution by the weightings in the objective function. That is, x_i contributes $1 + \lambda D_i^2$ to the objective, thus, we scale the individual components such that they contribute the same amount to the objective.

Exercise 4: Let $X \in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^m$ be given, and $\lambda > 0$. Consider the problem

$$w^* = \underset{w \in \mathbb{R}^n}{\arg\min} \|y - X^T w\|_2^2 + \lambda \|w\|_2^2$$

From the notes we know that there exists a unique solution w^* and that $w^* \in \mathcal{R}(X)$. Using the above and these two results, show that $w^* = X(X^T X + \lambda I_m)^{-1} y$.

Answer: We know that $w^* \in \mathcal{R}(X)$, so we have $Xa^* = w^*$ for some $a \in \mathbb{R}^m$. Thus, the objective becomes

$$a^* = \underset{a \in \mathbb{R}^m}{\arg\min} \|y - X^T X a\|_2^2 + \lambda \|Xa\|_2^2$$

Now, taking the gradient and setting to 0 yields

$$X^T X(X^T X a + \lambda I_m a - y) = \mathbf{0}$$

Since the solution is unique, we know that $(X^T X a + \lambda I_m a - y) \notin \mathcal{N}(X^T X)$, otherwise, any scaling would be optimal. Thus, we have $X^T X a + \lambda I_m a - y = 0$ which gives $a^* = (X^T X + \lambda I_m)^{-1} y$, where the matrix is invertible has been shown in the notes. Thus, we conclude that

$$w^* = Xa^* = X(X^T X + \lambda I_m)^{-1} y$$

Exercise 5: One form of regularized least squares can be posed as:

$$w^* = \underset{w \in \mathbb{R}^n}{\arg\min} \|Fw - y\|_2^2 + \lambda \|Gw - g\|_2^2$$

where $F \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, $G \in \mathbb{R}^{k \times n}$, $g \in \mathbb{R}^k$, and $\lambda > 0$.

a) Show that a sufficient condition for the above to have a unique solution is that $\text{rank}(G) = n$.

b) Show that a necessary and sufficient condition is that $\mathcal{N}(F) \cap \mathcal{N}(G) = \mathbf{0}$.

Answer:

a) We use the same reformulation of the problem and use

$$\|Fw - y\|_2^2 + \lambda \|Gw - g\|_2^2 = \|\tilde{F}w - \tilde{y}\|_2^2$$

where

$$\tilde{F} = \begin{bmatrix} F \\ \sqrt{\lambda} G \end{bmatrix} \in \mathbb{R}^{(m+k) \times n}, \text{ and } \tilde{y} = \begin{bmatrix} y \\ \sqrt{\lambda} g \end{bmatrix} \in \mathbb{R}^{m+k}$$

From this, it is evident that if $\text{rank}(G) = n$, then $\tilde{F}$ has rank of at least n as well as $\tilde{F}w = \begin{bmatrix} Fw \\ \sqrt{\lambda} Gw \end{bmatrix}$. Then, $\tilde{F}^T \tilde{F} \in \mathbb{R}^{n \times n}$ is full rank and hence the solution to the least squares problem is unique.

b) As before, we have that a unique solution exists if and only if $\tilde{F}^T \tilde{F}$ is invertible. But since $\mathcal{N}(F) \cap \mathcal{N}(G) = \mathbf{0}$, then we must have that $\mathcal{N}(\tilde{F}) = \mathbf{0}$ as $\tilde{F}w = \begin{bmatrix} Fw \\ \sqrt{\lambda} Gw \end{bmatrix}$ implies you cannot make both entries 0 simultaneously. Since $\mathcal{N}(\tilde{F}) = \mathbf{0}$, then $\tilde{F}$ is full rank and hence $\tilde{F}^T \tilde{F}$ is invertible. This proves the result.