Machine Learning 10-701 due October 17, 2013
http://alex.smola.org/teaching/cmu2013-10-701x/
Carnegie Mellon University

Homework 2 Solutions

---

1 Convexity [Dougal; 25 pts]

1.1 Calculus of convex functions

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^m$; define $h_1(x) = f(Ax + b)$. Then

$$\begin{aligned} h_1(\lambda x + (1-\lambda)y) &= f(A(\lambda x + (1-\lambda)y) + b) \\ &= f(\lambda Ax + (1-\lambda)Ay + b) \\ &= f(\lambda[Ax + b] + (1-\lambda)[Ay + b]) \\ &\leq \lambda f(Ax + b) + (1-\lambda) f(Ay + b) \\ &= \lambda h_1(x) + (1-\lambda) h_1(y). \end{aligned}$$

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$; define $h_2 = \max(f, g)$. Then

$$\begin{aligned} h_2(\lambda x + (1-\lambda)y) &= \max\left(f(\lambda x + (1-\lambda)y), g(\lambda x + (1-\lambda)y)\right) \\ &\leq \max\left(\lambda f(x) + (1-\lambda)f(y), \lambda g(x) + (1-\lambda)g(y)\right) \\ &\leq \max\left(\lambda f(x), \lambda g(x)\right) + \max\left((1-\lambda)f(y), (1-\lambda)g(y)\right) \\ &= \lambda \max\left(f(x), g(x)\right) + (1-\lambda)\max\left(f(y), g(y)\right) \\ &= \lambda h_2(x) + (1-\lambda) h_2(y). \end{aligned}$$

(c) Let $g : \mathbb{R} \to \mathbb{R}$ both convex and nondecreasing, $f : \mathbb{R}^n \to \mathbb{R}$ convex but not necessarily nondecreasing; define $h_3(x) = g(f(x))$. Then

$$\begin{aligned} h_3(\lambda x + (1-\lambda)y) &= g(f(\lambda x + (1-\lambda)y)) \\ &\leq g(\lambda f(x) + (1-\lambda)f(y)) \\ &\leq \lambda g(f(x)) + (1-\lambda) g(f(y)) \\ &= \lambda h_3(x) + (1-\lambda) h_3(y). \end{aligned}$$

1.2 First-order condition

Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and $\operatorname{dom} f$ be open.

Suppose that f is convex. For any x and y in its domain, $(x, f(x))$ and $(y, f(y))$ are in the epigraph; then $(x + \lambda(y - x), f(x) + \lambda(f(y) - f(x)))$ is also in the epigraph for any $\lambda \in [0, 1]$. Thus $x + \lambda(y - x) \in \operatorname{dom} f$, so $\operatorname{dom} f$ must be convex. We also have that

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

Taking the limit as $\lambda(y - x) \to 0$ from above, we get $f'(x)\,(y - x) \leq f(y) - f(x)$ as desired.

Suppose that $\operatorname{dom} f$ is convex and $f(b) - f(a) \geq f'(a)\,(b - a)$ for all points a, b. Then for any $\lambda \in [0, 1]$, $x, y \in \operatorname{dom} f$, $z = \lambda x + (1 - \lambda)y \in \operatorname{dom} f$. Then $f(x) - f(z) \geq f'(z)(x - z)$ and $f(y) - f(z) \geq f'(z)(y - z)$.

Multiplying the first inequality by λ and the second by $(1-\lambda)$:

$$\lambda(f(x)-f(z))+(1-\lambda)(f(y)-f(z)) \geq \lambda f'(z)(x-z)+(1-\lambda)(f'(z)(y-z))$$
$$\lambda f(x)-\lambda f(z)+f(y)-f(z)-\lambda f(y)+\lambda f(z) \geq f'(z)\left[\lambda x-\lambda z+y-z-\lambda y+\lambda z\right]$$
$$\lambda f(x)+(1-\lambda)f(y)-f(z) \geq f'(z)\left[\lambda x+(1-\lambda)y-z\right]=0$$

by the definition of z. Thus $\lambda f(x)+(1-\lambda)f(y) \geq f(x+(1-\lambda)f(y))$, and so $(\lambda f(x)+(1-\lambda)f(y), \lambda f(x)+(1-\lambda)f(y))$ is in the epigraph of f. Since this is true for all $\lambda \in [0,1]$, the epigraph of f must be convex.

1.3 Strict and strong convexity

(a) Let f be an m-strongly convex function. By definition, for any $x, y \in \operatorname{dom} f, \lambda \in [0,1]$:

$$\begin{aligned} f(\lambda x+(1-\lambda)y) &\leq \lambda f(x)+(1-\lambda)f(y)-\frac{1}{2}m\lambda(1-\lambda)\|x-y\|_2^2 \\ &= \lambda f(x)+(1-\lambda)f(y)-\frac{1}{2}n\lambda(1-\lambda)\|x-y\|_2^2+\frac{1}{2}(n-m)\lambda(1-\lambda)\|x-y\|_2^2 \\ &\leq \lambda f(x)+(1-\lambda)f(y)-\frac{1}{2}n\lambda(1-\lambda)\|x-y\|_2^2 \end{aligned}$$

since $n-m<0$, and λ, $1-\lambda$, and $\|x-y\|_2^2$ are all nonnegative.

(b) Let f be an m-strongly convex function. By definition, for any $x \neq y \in \operatorname{dom} f, \lambda \in (0,1)$:

$$\begin{aligned} f(\lambda x+(1-\lambda)y) &\leq \lambda f(x)+(1-\lambda)f(y)-\frac{1}{2}m\lambda(1-\lambda)\|x-y\|_2^2 \\ &< \lambda f(x)+(1-\lambda)f(y) \end{aligned}$$

since m, λ, $1-\lambda$, and $\|x-y\|_2^2$ are all positive.

(c) One solution is $f(x)=e^x$.

- Note that $f''(x)=f'(x)=f(x)$, so that $\nabla^2 f(x)=e^x>0$ for all x, and by the second-order condition f is strictly convex.
- But f is not m-strongly convex for any m. For that to be true, there would have to be some $m>0$ for which $f''(x)>m$ for all x. But then we'd have $f''(\log m-1)=e^{\log m-1}=\frac{1}{e}m<m$, a contradiction.

Another possible solution is $f(x)=x^4$, a case where we actually have $f''(0)=0$. Then:

- $f(x)$ is not m-strongly convex for any m. If it were, there would be an m such that $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbb{R}$, since f is twice differentiable. But $\nabla^2 f(x)=12x^2$, which means $\nabla^2 f(0)=0$.
- $f(x)$ is strictly convex. There may be a nicer proof, but we will verify the first-order condition

$$(\lambda x+(1-\lambda)y)^4<\lambda x^4+(1-\lambda)y^4 \tag{1}$$

for all $\lambda \in (0,1)$, $x \neq y \in \mathbb{R}$.

  - First, we can see that x^4 is $(12\varepsilon^2)$-strongly convex on (ε,∞). Thus, by part (b), (1) holds for all $x>0$, $y>0$.
  - x^4 is also $(12\varepsilon^2)$-strongly convex on $(-\infty,-\varepsilon)$. So (1) holds for all $x<0$, $y<0$.

Homework 2 Solutions

- Suppose $x > 0$, $y < 0$, $|x| \neq |y|$. Note

  $$(\lambda x + (1-\lambda)y)^4 = \lambda^4 x^4 + 4\lambda^3 x^3 (1-\lambda)y + 6\lambda^2 x^2 (1-\lambda)^2 y^2 + 4\lambda x (1-\lambda)^3 y^3 + (1-\lambda)^4 y^4.$$

  Since (1) holds for x and $|y|$, we know that

  $$\lambda^4 x^4 + 4\lambda^3 x^3 (1-\lambda)|y| + 6\lambda^2 x^2 (1-\lambda)^2 |y|^2 + 4\lambda x (1-\lambda)^3 |y|^3 + (1-\lambda)^4 |y|^4 < \lambda x^4 + (1-\lambda)|y|^4.$$

  But $|y| = -y$, $|y|^2 = y^2$, $|y|^3 = -y^3$, $|y| = y^4$, so we have

  $$\lambda^4 x^4 - 4\lambda^3 x^3 (1-\lambda)y + 6\lambda^2 x^2 (1-\lambda)^2 y^2 - 4\lambda x (1-\lambda)^3 y^3 + (1-\lambda)^4 y^4 < \lambda x^4 + (1-\lambda)y^4.$$

  Note that $\lambda^3 x^3 (1-\lambda)y < 0$, so we can add 8 times that to the LHS without breaking the inequality. The same is true for $\lambda x (1-\lambda)^3 y^3$. We then get

  $$\lambda^4 x^4 + 4\lambda^3 x^3 (1-\lambda)y + 6\lambda^2 x^2 (1-\lambda)^2 y^2 + 4\lambda x (1-\lambda)^3 y^3 + (1-\lambda)^4 y^4 < \lambda x^4 + (1-\lambda)y^4$$

  as desired.
- Suppose $x < 0$, $y > 0$, $|x| \neq |y|$. By symmetry with the last part, (1) holds.
- Suppose $y = -x$. Then

  $$(\lambda x + (1-\lambda)y)^4 = (\lambda x - (1-\lambda)x)^4 = (2\lambda - 1)^4 x^4$$
  $$\lambda x^4 + (1-\lambda)y^4 = \lambda x^4 + (1-\lambda)x^4 = x^4.$$

  Since $0 < \lambda < 1$, $-1 < 2\lambda - 1 < 1$. Thus $(2\lambda-1)^4 < 1$, and $(2\lambda-1)^4 x^4 < x^4$, and (1) holds.

We have thus shown that (1) holds for all $x, y \in \mathbb{R}$, so that x^4 is strictly convex.

1.4 Examples

(a) The second derivative of $x^2 + x^4$ is $2 + 12x^2 \geq 2$, so $x^2 + x^4$ is 2-strongly convex on $\mathbb{R}$.

(b) $x^2 + x^4$ is still strongly-convex on $[1, 5]$. It's 14-strongly convex, in fact, though we didn't ask for the constant.

(c) An arbitrary norm is convex, because $\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|$. It is not necessarily strictly convex; a simple counterexample is the absolute value on $\mathbb{R}$, where if $x > 0$, $y > 0$ we have $|\lambda x + (1-\lambda)y| = \lambda x + (1-\lambda)y$.

2 Linear Regression, Again ? [Ahmed; 20 pts]

2.1 Why Lasso Works

(a)

$$\begin{aligned}
J_\lambda(\beta) &= \frac{1}{2}\|y - X\beta\|^2 + \lambda|\beta|_1 \\
&= \frac{1}{2}(\|y\|^2 + \beta^T X^T X\beta - 2y^T X\beta) + \lambda|\beta|_1 \\
&= \frac{1}{2}(\|y\|^2 + \|\beta\|^2 - 2y^T X\beta) + \lambda|\beta|_1 \\
&= \frac{1}{2}\|y\|^2 + \sum_{i=1}^{d} \left(\frac{1}{2}\beta_i^2 - y^T X_{.i}\beta_i + \lambda|\beta_i|\right)
\end{aligned}$$

Homework 2 Solutions

(b) Note that $d|\beta|/d\beta = 1$ iff $\beta > 0$. Setting the partial derivative of the objective function w.r.t β_j to 0 we get

$$\frac{\partial}{\partial \beta_j} J_\lambda(\beta) = \frac{\partial}{\partial \beta_j} f(X_{.j}, y, \beta_j, \lambda) = \beta_j - y^T X_{.j} + \lambda = 0$$

, which gives

$$\beta_j^* = y^T X_{.j} - \lambda$$

(c) Note that $d|\beta|/d\beta = 1$ iff $\beta < 0$. Using the same procedure we can show that

$$\beta_j^* = y^T X_{.j} + \lambda$$

(d) $\beta_j^* = 0$ what it can neither be greater then or less than 0– that is, when

$$y^T X_{.j} - \lambda < 0,$$
$$y^T X_{.j} + \lambda > 0$$

which can be formulated as

$$|y^T X_{.j}| < \lambda$$

Note that $y^T X_{.j}$ indicates how much $X_{.j}$ and y are (anti)correlated– that is, how strong $X_{.j}$ is as a predictor for y. This condition means that β_j^* will be set to 0 if the corresponding feature is not (anti)correlated enough with the output.

(e) Setting the partial derivative of the objective function w.r.t β_j to 0 we get

$$\beta_j - y^T X_{.j} + \lambda \beta_j = 0$$

which means $\beta_j^* = 0$ iff $y^T X_{.j}$ is exactly 0. This is a much stronger condition than the lasso case.

2.2 Kernel Ridge Regression

(a) One way to show it is to write β^* as $X^T c$ for some vector c:

$$(X^T X + \lambda I)\beta^* = X^T y$$
$$\beta^* = \lambda^{-1}(X^T y - X^T X \beta^*) = X^T(\lambda^{-1}(y - X\beta^*)) = X^T c,$$

where

$$c = \lambda^{-1}(y - X\beta^*)$$

Another way is to use the orthogonal decomposition $\beta = \beta_\parallel + \beta_\perp$ where $\beta_\perp$ is the component orthogonal to all training points. Then $X\beta_\perp = 0$ and we get

$$J(\beta) = \frac{1}{2}\|y - X\beta_\parallel - X\beta_\perp\|^2 + \frac{1}{2}\|\beta_\parallel\|^2 + \frac{1}{2}\|\beta_\perp\|^2 \geq J(\beta_\parallel),$$

with equality holding only if $\beta_\perp = 0$, which means that unless $\beta_\perp = 0$, β cannot be optimal.

Homework 2 Solutions

(b) Note that $\beta = X^T\alpha$

$$\begin{aligned}(X^TX + \lambda I)\beta^* &= X^Ty \\ (X^TX + \lambda I)X^T\alpha^* &= X^Ty \\ X^TXX^T\alpha^* + \lambda X^T\alpha^* &= X^Ty \\ X^T(XX^T + \lambda I)\alpha^* &= X^Ty\end{aligned}$$

The last equality shown α^* given by

$$(XX^T + \lambda I)\alpha^* = y,$$

results in the optimal β^*, which is the desired result. The part that depends on training inputs is XX^T, but $(XX^T)_{i,j} = \langle x_i, x_j \rangle$

(c)

$$\hat{f}(x) = \beta^T x = \sum_i \alpha_i x_i^T x = \sum_i \alpha_i \langle x_i, x \rangle$$

(d) For non-kernelized version we need d numbers to store β, for the kernelized version we need n numbers to store α and $n \times d$ numbers to store training inputs.