

10-702 Statistical Machine Learning: Assignment 2

Due Friday, February 5

Hand in to Sharon Cavlovich, GHC (Gates Hillman Center) 8215 by 3:00. Use R for all numerical computations.

1. (Convex sets and functions)

(a) For $x \in \mathbb{R}^n$ define the ℓ_p norm

$$\|x\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}$$

for $p > 0$. Let

$$C = \left\{ x : \|x\|_p \leq 1 \right\}.$$

Show that C is convex set if and only if $p \geq 1$.

(b) Check if the following functions are convex, concave, or neither convex nor concave:

i.

$$f(x, y, z) = \max \left(1 + |x| - y, \frac{1}{\sqrt{z}}, 0 \right) + 2 \cdot \frac{(x - z)^2}{y + 1}$$

with $y + 1 > 0$ and $z > 0$.

ii. $f(X) = (\det(X))^{1/n}$ where $X \in \mathbf{S}_{++}^n$

(c) If $C \subset \mathbb{R}^n$ is an open convex set, a function $G : C \mapsto \mathbb{R}^n$ is a *monotone mapping* on C if $(G(x) - G(y))^T(x - y) \geq 0$ for all $x, y \in C$. Show that for a differentiable, convex function $f : C \mapsto \mathbb{R}$, the gradient ∇f is a monotone mapping.

2. (Convex conjugacy)

(a) If $f(x, y) = f_1(x) + f_2(y)$, with f_1 and f_2 convex, show that

$$f^*(x, y) = f_1^*(x) + f_2^*(y).$$

Does this hold if f_1 and f_2 are not convex?

(b) Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex mapping and let $\lambda_1, \dots, \lambda_n > 0$ be positive weights such that $\sum_{i=1}^n \lambda_i = 1$. For a sequence of numbers $x_1, \dots, x_n \in \mathbb{R}$, use conjugate duality to prove the inequality

$$f \left(\sum_{i=1}^n \lambda_i x_i \right) \leq \sum_{i=1}^n \lambda_i f(x_i). \quad (1)$$

(c) For each of the following functions

- Show that f is convex
- Calculate the conjugate dual f^* .
- Calculate the bi-conjugate f^{**} , and determine if it is equal to f .
- Plot f, f^* and f^{**} .

i. $f(x) = x \ln(x) - x$ for $x > 0$ and ∞ otherwise.

ii. $f(x) = \sup_{y \in C} y^T x$, where C is a non-empty, closed and convex set.

iii. $f(x) = [\max(1 - x, 0)]^2$ for $x \in \mathbb{R}$.

iv. $f(x) = \max(1 - x, 0)$ for $x \in \mathbb{R}$.

(d) Let $f(A)$ be the function defined on the set of positive-definite $p \times p$ matrices as

$$f(A) = \text{tr}(AS) - \log \det A$$

where S is a given matrix.

- i. Derive the conjugate f^* .
- ii. Derive the dual of the primal problem

$$\min_{A \in \mathcal{S}_{++}^p} f(A) \tag{2}$$

$$\text{such that } \text{tr}(A) \leq L \tag{3}$$

using the conjugate f^* derived in the previous part.

3. (Subdifferentials and thresholding)

(a) Let $Z \sim \mathcal{N}(\mu, \sigma^2)$ with known variance σ^2 . Consider the following optimization problem:

$$\hat{\mu} = \arg \min_{x \in \mathbb{R}} (Z - x)^2 + \lambda \text{pen}(x)$$

where $\lambda \geq 0$ and $\text{pen} : \mathbb{R} \mapsto [0, \infty)$ is the penalty function defined as follows:

- i. $\text{pen}(x) = |x|$
- ii. $\text{pen}(x) = x^2$
- iii. $\text{pen}(x) = \mathbb{I}\{x \neq 0\}$.

For each of the penalty functions find an explicit expression for $\hat{\mu}$. We would like you to derive these results directly, without referring to equations and theorems in the handouts from class.

The penalty $\text{pen}(x) = x^2$ corresponds to the ridge regression penalty; in class we learned that it does not encourage sparsity. However, assume that the parameter λ is chosen as $\lambda = \frac{\lambda'}{|\mu|}$, for some other $\lambda' \geq 0$. For this choice of the parameter λ ,

is the solution going to be sparse? What is the correspondence to the Lasso penalty $\text{pen}(x) = |x|$?

Hint: You may want to compute the probability $\mathbb{P}[\hat{\mu} = 0 \wedge \mu \neq 0]$ and $\mathbb{P}[\hat{\mu} \neq 0 \wedge \mu = 0]$ for the two different penalties.

- (b) Let $Z \sim \mathcal{N}(\mu, \sigma^2 I_p)$ be distributed according to a multivariate Gaussian distribution with known σ^2 . Assume that the components of the mean vector μ are naturally grouped as

$$\begin{aligned} \mu &= (\underbrace{\mu_{1,1}, \dots, \mu_{1,p_1}}_{\mathcal{G}_1}, \dots, \underbrace{\mu_{j,1}, \dots, \mu_{j,p_j}}_{\mathcal{G}_j}, \dots, \underbrace{\mu_{G,1}, \dots, \mu_{G,p_G}}_{\mathcal{G}_G}) \\ &= (\mu_{\mathcal{G}_1}, \dots, \mu_{\mathcal{G}_j}, \dots, \mu_{\mathcal{G}_G}) \end{aligned}$$

with $\mu_{\mathcal{G}_j} \in \mathbb{R}^{p_j}$ and $\sum_{j=1}^G p_j = p$. Consider the optimization problem

$$\hat{\mu} = \arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \sum_{j=1}^G \|Z_{\mathcal{G}_j} - x_{\mathcal{G}_j}\|^2 + \lambda_1 \sum_{j=1}^G \sum_{i=1}^{p_j} |x_{j,i}| + \lambda_2 \sum_{j=1}^G \|x_{\mathcal{G}_j}\|_2.$$

Find an explicit expression for $\hat{\mu}$. (Remark: You may use any results from the class notes for this part of the problem.)

- (c) Let $\Psi(z) = \|z\|_q$ with $z \in \mathbb{R}^p$ and let $\Psi^*(\cdot)$ denote the conjugate. Prove that

$$w \in \partial\Psi(y) \quad \text{if and only if} \quad y \in \partial\Psi^*(w),$$

where $\partial\Psi(z)$ denotes the subdifferential of Ψ at z .

4. (a) Generate data as follows.

```
n      = 100
p      = 1000
X      = matrix(rnorm(n*p), n, p)
beta   = c(rep(10, 5), rep(5, 5), rep(0, 990))
y      = X %*% beta + rnorm(n)
```

- (b) Write an R function to do marginal regression, which is described in Chapter 11 (Model Selection) Section 7, and apply it to the generated data. By varying t from a large value to a small one, you can obtain a sequence of models. Output the cross-validation score (5-fold) for the sequence.
- (c) Write an R function to do forward stepwise regression. Output the cross-validation score for the sequence of models.
- (d) Run the lasso on these data and output the cross-validation score. You can use the package `lars` for this:

```
install.packages("lars") ### only do this once
library(lars)
help(lars)
```

(e) Compare the methods and summarize your results.