CX4240 Computing for Data Analysis - Homework 2

Name:
GTID:

Deadline: 11:59 pm EST, Feb 23

- Submit your answers as one single PDF file on Gradescope.
- You will be allowed 2 total late days (48 hours) without penalty for the entire semester. Once those days are used, you will be penalized according to the following policy:
  - Homework is worth full credit before the due time.
  - It is worth 75% credit for the next 24 hours.
  - It is worth 50% credit for the second 24 hours.
  - It is worth zero credit after that.
- You are required to use Latex, or word processing software, to generate your solutions to the written questions. Handwritten solutions WILL NOT BE ACCEPTED.

1 Linear Regression [45 pts]

Let $\{(x_i, y_i)\}_{i=1}^n$ be a dataset where $x_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \mathbb{R}$ is the target value. Consider the linear regression model with Gaussian noise:

$$y_i = \theta^\top x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $\theta \in \mathbb{R}^d$ and $\sigma^2 > 0$ is known.

Define the design matrix $X \in \mathbb{R}^{n \times d}$ whose i-th row is $x_i^\top$, and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.

1. **[15 points] From Gaussian likelihood to least squares**

   (a) Write the likelihood $L(\theta) = p(y \mid X, \theta)$ under the Gaussian noise assumption, and simplify the log-likelihood
   $$\ell(\theta) = \log L(\theta)$$
   up to additive constants independent of θ.

   (b) Prove that maximizing $\ell(\theta)$ over θ is equivalent to minimizing the squared error:
   $$\|y - X\theta\|_2^2.$$
   (Hint: You may ignore additive/multiplicative constants that do not change the optimizer.)

(c) Compute the gradient $\nabla_\theta \|y - X\theta\|_2^2$, and show that any stationary point (i.e., a point where the gradient equals zero) satisfies the normal equation

$$X^\top X\theta = X^\top y.$$

2. **[30 points] MAP estimation and ridge regression**

Now place a Gaussian prior on θ:

$$\theta \sim \mathcal{N}(0, \tau^2 I_d),$$

where $\tau^2 > 0$ is known.

(a) Using Bayes' rule, write the (unnormalized) log-posterior

$$\log p(\theta \mid X, y)$$

up to additive constants independent of θ.

(b) Prove that the MAP estimator

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \log p(\theta \mid X, y)$$

is equivalent to the ridge regression optimization problem:

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \left(\|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right),$$

for an appropriate $\lambda > 0$ expressed in terms of σ^2, τ^2.
(You must explicitly give λ as a function of σ^2, τ^2.)

(c) Define the ridge objective:

$$J(\theta) = \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2.$$

Compute its gradient $\nabla_\theta J(\theta)$ and show that any minimizer (in particular $\hat{\theta}_{\text{MAP}}$) satisfies

$$(X^\top X + \lambda I_d)\hat{\theta}_{\text{MAP}} = X^\top y.$$

Solution:

.

1. **From Gaussian likelihood to least squares.**

(a) **Write the likelihood and log-likelihood.**

Under the model

$$y_i = \theta^\top x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

the conditional density of y_i is

$$p(y_i \mid x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \theta^\top x_i)^2}{2\sigma^2} \right).$$

Since the samples are independent, the likelihood is

$$L(\theta) = p(y \mid X, \theta) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 \right).$$

Taking logarithms,

$$\ell(\theta) = \log L(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2.$$

Up to additive constants independent of θ,

$$\ell(\theta) \equiv -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2.$$

(b) **Show equivalence to least squares.**
Since $\sigma^2 > 0$ is fixed, maximizing $\ell(\theta)$ is equivalent to minimizing

$$\sum_{i=1}^{n} (y_i - \theta^\top x_i)^2.$$

Let $r(\theta) = y - X\theta$ be the residual vector. Then

$$\sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 = \|y - X\theta\|_2^2.$$

Hence,

$$\arg\max_\theta \ell(\theta) = \arg\min_\theta \|y - X\theta\|_2^2.$$

(c) **Compute the gradient and obtain the normal equation.**
Expand the squared norm:

$$\|y - X\theta\|_2^2 = (y - X\theta)^\top (y - X\theta) = y^\top y - 2\theta^\top X^\top y + \theta^\top X^\top X\theta.$$

Differentiate with respect to θ:

$$\nabla_\theta \|y - X\theta\|_2^2 = -2X^\top y + 2X^\top X\theta = 2X^\top (X\theta - y).$$

Setting the gradient to zero gives

$$X^\top (X\theta - y) = 0 \quad \Longrightarrow \quad X^\top X\,\theta = X^\top y,$$

which is the normal equation.

2. **MAP estimation and ridge regression.**

(a) **Write the log-posterior up to constants.**
Assume the prior

$$\theta \sim \mathcal{N}(0, \tau^2 I_d),$$

so

$$p(\theta) = (2\pi\tau^2)^{-d/2} \exp\left(-\frac{1}{2\tau^2}\|\theta\|_2^2\right).$$

By Bayes' rule,

$$p(\theta \mid X, y) \propto p(y \mid X, \theta)\, p(\theta).$$

Taking logs and ignoring constants independent of θ,

$$\log p(\theta \mid X, y) \equiv -\frac{1}{2\sigma^2}\|y - X\theta\|_2^2 - \frac{1}{2\tau^2}\|\theta\|_2^2.$$

(b) **Show MAP is equivalent to ridge regression and identify λ.**
The MAP estimator maximizes the log-posterior, equivalently minimizes its negative:

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \left(\frac{1}{2\sigma^2}\|y - X\theta\|_2^2 + \frac{1}{2\tau^2}\|\theta\|_2^2\right).$$

Multiplying the objective by $2\sigma^2$ (which does not change the minimizer),

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \left(\|y - X\theta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\theta\|_2^2\right).$$

Thus the ridge regression form holds with

$$\lambda = \frac{\sigma^2}{\tau^2} > 0.$$

(c) **Compute the gradient, obtain the closed form, and prove uniqueness.**
Define the ridge objective:

$$J(\theta) = \|y - X\theta\|_2^2 + \lambda\|\theta\|_2^2.$$

Its gradient is

$$\nabla_\theta J(\theta) = 2X^\top(X\theta - y) + 2\lambda\theta.$$

Setting the gradient to zero:

$$2X^\top X\theta - 2X^\top y + 2\lambda\theta = 0 \quad \Longleftrightarrow \quad (X^\top X + \lambda I_d)\theta = X^\top y.$$

2 Logistic Regression [55 pts]

In this exercise, you will implement logistic regression to predict a binary label from input features. We have prepared a Jupyter notebook to guide you through the complete logistic regression workflow.

Please access the notebook using this link. You can also download the notebook by using `File` (top-left corner) → `Download` → `Download .ipynb`.

For this problem, execute all cells in the notebook to generate the required outputs, and export the notebook into a single PDF file. For example, if you are using Google Colab, you can use `File` (top-left corner) → `Print` to generate a PDF. For submission, a separate homework entry for this problem will be created on GradeScope. Please upload the generated PDF to that entry.

Solution:
Solution Notebook Link

References

Solution:
Please mention any AI tools, people, post or blog etc. you used.