BIOENG-210: Biological Data Science I: Statistical Learning

Theoretical Exercise Week 9
Prof. Gioele La Manno

April 2024

1 MLE and MAP for Linear Models

In all of the following parts, write your answer as the solution to a norm minimization problem, potentially with a regularization term. **You do not need to solve the optimization problem.** Simplify any sums using matrix notation for full credit.

Hint: Recall that the MAP estimator maximizes $P(\boldsymbol{\theta}|Y)$.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} P(Y|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

The difference between MAP and MLE is the inclusion of a prior distribution on $\boldsymbol{\theta}$ in the objective function.

For the following problems assume you are given $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ as your data.

(a) Let $y = X\boldsymbol{\theta} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Sigma)$ for some positive definite, diagonal Σ. Write the MLE estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares problem, potentially with a regularization term.

(b) Let $y|\boldsymbol{\theta} \sim \mathcal{N}(X\boldsymbol{\theta}, \Sigma)$ for some positive definite, diagonal Σ. Let $\boldsymbol{\theta} \sim \mathcal{N}(0, \lambda I_d)$ for some $\lambda > 0$ be the prior on $\boldsymbol{\theta}$. Write the MAP estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

(c) Let $y = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\epsilon_i \overset{i.i.d.}{\sim} \text{Laplace}(0, 1)$. Recall that the pdf for $\text{Laplace}(\mu, b)$ is $p(x) = \frac{1}{2b}\exp\left(-\frac{1}{b}|x-\mu|\right)$. Write down the MLE estimator of $\boldsymbol{\theta}$ as the solution to a norm minimization optimization problem.

(d) Let $y|\boldsymbol{\theta} \sim \mathcal{N}(X\boldsymbol{\theta}, \Sigma)$ for some positive definite, diagonal Σ. Let $\theta_i \overset{i.i.d.}{\sim} \text{Laplace}(0, \lambda)$ for some positive scalar λ. Write the MAP estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

2 Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be independent samples from the following distribution:

$$P(x \mid \theta) = \theta x^{-\theta-1} \quad \text{where } \theta > 1,\, x \geq 1$$

Find the maximum likelihood estimator of θ.

3 Linear models and linear transformation

In this exercise we are going to see how the solution to the least squares problem changes when a linear transformation is applied to the input features X. Recall that in linear regression given $X \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ we aim to find the set of coefficients $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ that minimizes:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\hat{\boldsymbol{\beta}}} ||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}||_2^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} X_{ij}\beta_j)^2 \tag{1}$$

For simplicity, we can define $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$. Recall that the solution is given by:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{2}$$

For all the exercise, assume that X is full rank and therefore $X^T X$ is invertible and also $n > d$. We would like to linearly transform our features, so that we obtain a new set of features $X' \in \mathbb{R}^{n \times d'}$. If our the matrix defining our linear transformation is $A \in \mathbb{R}^{d \times d'}$, the transformed features X are simply given by:

$$X' = XA \tag{3}$$

Now we would like to find the set of coefficients $\hat{\boldsymbol{\beta}}'$ that minimize:

$$\hat{\boldsymbol{\beta}}' = \text{argmin}_{\hat{\boldsymbol{\beta}}'} ||\boldsymbol{y} - X'\hat{\boldsymbol{\beta}}'||_2^2 \tag{4}$$

a) Write down the solution to 4, that is, what is the optimal $\hat{\boldsymbol{\beta}}'$ in terms of X' and $\boldsymbol{y}$.

   We will now try to relate this solution to 2.

b) Substitute $X' = XA$ to the expression found for $\hat{\boldsymbol{\beta}}'$. You will not be able to simply much. Hint: Remember that given two matrices A, B, $(AB)^T = B^T A^T$.

   From now on, assume that $d = d'$ and that A is full rank (thus invertible). This assumption is equivalent to saying that we transform the data "without loss of information".

c) Show that $\hat{\boldsymbol{\beta}}' = A^{-1}\hat{\boldsymbol{\beta}}$ Hint: The same property as before also holds for the inverse $(AB)^{-1} = B^{-1}A^{-1}$ if A and B are full rank and squared.

d) Show that the predictions of the model do not change if we fitted with the transformed data X'.

e) In part c, we have assumed $d' = d$. What would happen to the solution to the least squares problem in the case $d' > d$? Hint: $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$

f) Finally, what would you intuitively think happens in the case $d' < d$. Try to reason in terms of model performance when comparing the model fitted with X and XA (with $d' < d$).

4 Statistical Properties of the Uniform Distribution

Consider a continuous uniform distribution defined on the interval $[a, b]$ with length $L = b - a$.

1. Derive the probability density function (PDF) of this uniform distribution.
2. Calculate the expectation value (mean) of this distribution and express it as a function of L and a.
3. Calculate the variance of this distribution and express it as a function of L only.

Hint: Remember that for a continuous random variable X with probability density function $f(x)$:

- The expectation is given by $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$
- The variance is given by $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

5 James-Stein estimator

Problem 1: Setup the Multivariate Normal Model
Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ where each $X_i \sim N(\theta_i, \sigma^2)$ independently.

a) What is the MLE for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$? (**Hint**: Note that each variable X_i has a different mean.)

b) Show that the risk (mean squared error) of the MLE, $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] = p\sigma^2$, where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

Problem 2: Introduce the James-Stein Estimator
Define a James-Stein estimator:

$$\hat{\boldsymbol{\theta}}^{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X},$$

where $\|\boldsymbol{X}\|^2 = \sum_{i=1}^{p} X_i^2$. Compute the condition that p should satisfy so that the shrinkage factor is positive.

Problem 3: Classical vs. Shrinkage Estimators

a) Mention the trade-off we make in terms of bias and variance between the JS estimator and the MLE.

b) Explain the importance of the James-Stein estimator in practical applications. Where might we expect it to outperform traditional methods, and why?

6 Pen-and-Paper PCA Exercise

Exercise

Consider the following dataset of four observations in two dimensions:

Obs.	x	y
1	1	2
2	2	1
3	3	4
4	4	3

1. Compute the sample means $\bar{x}$ and $\bar{y}$.
2. Center the data by subtracting $(\bar{x}, \bar{y})$ from each point.
3. Form the sample covariance matrix

$$S = \frac{1}{n-1} \sum_{i=1}^{4} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix} (x_i - \bar{x},\ y_i - \bar{y}).$$

4. Solve for the eigenvalues λ_1, λ_2 of S.
5. Find corresponding (unit) eigenvectors $v^{(1)}, v^{(2)}$.
6. Compute the proportion of total variance explained by each principal component.
7. Project each centered point onto the first principal component.
8. Sketch the centered data, overlay the PC axes, and draw the ellipse with semi-axes $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.