BIOENG-210: Biological Data Science I: Statistical Learning

Theoretical Exercise Week 11
Prof. Gioele La Manno

May 2024

1 [OPTIONAL]Uniqueness of the solution to the linear regression problem

Note: Although this exercise is mathematically more demanding, we do not expect you to solve this sort of problems in the exam. This is to illustrate a result that you should know of and make you feel more comfortable with the mathematical tools often used in data science and machine learning.

In this exercise we are going to discuss whether the optimal solutions of the different linear models are unique or not. Remember that the optimal parameters are always found by minimizing the negative log-likeihood, therefore we need to show that the minimum we find is a global minimum of the function we are optimizing.

We will start by showing it in the case of linear regression, which you might recall it consists of solving the optimization problem:

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 = \min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) \tag{1}$$

a) Start by computing the first derivative of Equation 1, $\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$. Hint: A couple of vector calculus identities that you can use:

   - $\nabla_{\mathbf{x}} ||\mathbf{x}||_2^2 = \nabla_{\mathbf{x}} \mathbf{x}^t \mathbf{x} = 2\mathbf{x}$
   - $\nabla_{\mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}^T$

b) Now compute the second derivative (or Hessian) with respect to the parameters ($\mathbf{H} = \nabla_{\boldsymbol{\beta}}(\nabla_{\boldsymbol{\beta}})\mathcal{L}(\boldsymbol{\beta})$) Hint: Recall that for any matrix $\mathbf{A}$, $(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T\mathbf{A}$

   You should have obtained a second derivative is constant for all values of $\boldsymbol{\beta}$.

Since $\mathcal{L}(\boldsymbol{\beta})$ is continuous and differentiable, in order to show that the solution is unique, we only need to show that the function $\mathcal{L}(\boldsymbol{\beta})$ is convex. This means that there are no changes in curvature in the function and that the minimum we find has to be the only one. To visualize this, imagine a function in 1D that is always convex (U-shapes) at every point, you should see there can only be one minimum, since the curvature does not change. For functions in $\mathbb{R}^n$, showing that a function is convex is equivalent to showing that the Hessian is positive definite ($\mathbf{H} \succ 0$), which translates to the following condition:

$$\mathbf{H} \succ 0 \leftrightarrow \mathbf{v}^T H \mathbf{v} > 0, \forall \mathbf{v} \neq \mathbf{0}$$

c) Show that the Hessian is positive definite and therefore that the solution to the linear regression problem is unique (assume $\mathbf{X}$ is full rank).

d) We will now repeat the same steps for the case of logistic regression, start by computing the first derivative. Recall that the optimization problem we are solving is:

$$\min_{\boldsymbol{\beta}} - \sum_{i=1}^{n} y_i \log(\sigma(\mathbf{x}_i^T \boldsymbol{\beta})) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})) \tag{2}$$

Hint: You can use that $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$ and recall that $\sigma(x) > 0, \forall x$.

e) As before, compute the second derivative (or Hessian) with respect to the parameters.

f) Finally, show that the Hessian is positive definite and therefore that the solution to the logistic regression problem is unique (assume $\mathbf{X}$ is full rank). Hint: You can use the fact that $\sigma(x)(1 - \sigma(x)) \geq 0, \forall x$.

2 Clustering

In this exercise, you are analyzing a dataset consisting of **500 single-cell gene expression profiles**, each measured across **1000 genes**.
After performing **Principal Component Analysis (PCA)**, you retain the top **3 principal components**, which capture 85% of the total variance in the data.

You wish to identify subpopulations of cells using unsupervised clustering.

2.1 Clustering Theory

(a) Briefly explain two limitations of **K-means clustering** when applied to high-dimensional biological data.
Discuss assumptions about cluster shapes, distance metrics, or initialization sensitivity.

(b) Define the **within-cluster sum of squares (WCSS)** objective function used by K-means. Show how this cost is minimized in the centroid update step.

(c) Based on the figure below, explain and justify which choice of k is optimal for the K-means clustering algorithm.

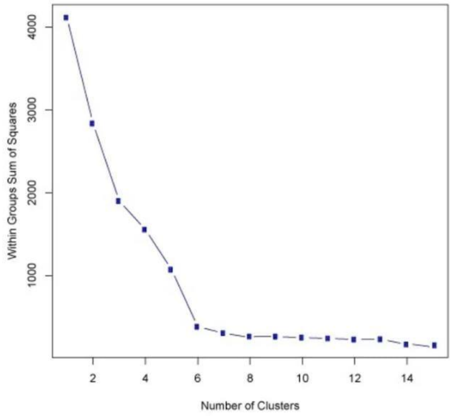

Figure 1: WCSS vs. number of clusters

2.2 Clustering Computation

You perform K-means clustering on the PCA-reduced data with $k = 3$, resulting in the following cluster centroids:

$$\mu_1 = [1.2,\ -0.5,\ 0.3]$$
$$\mu_2 = [3.5,\ 1.1,\ -0.8]$$
$$\mu_3 = [-0.9,\ -2.0,\ 1.5]$$

You are given three new cell profiles projected into PCA space:

$$\mathbf{x}_A = [1.0,\ -0.6,\ 0.2]$$
$$\mathbf{x}_B = [3.2,\ 1.4,\ -0.6]$$
$$\mathbf{x}_C = [-1.0,\ -1.8,\ 1.3]$$

(a) Assign each of the three cells (A, B, C) to a cluster using the **Euclidean distance**. Show all your calculations.

(b) Compute the **silhouette coefficient** $s(i)$ for **Cell A**, given:

$$a(i) = 0.45, \quad b(i) = 2.05$$

Use the formula:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i),\ b(i))}$$

(c) Suppose that running K-means with $k = 4$ yields a drop in the average silhouette score from **0.71** to **0.55**. Interpret this result. What does this suggest about your choice of k? What alternative method could be used to select an optimal k?

3 An almost-1-D cloud

Data set. Consider the six two-dimensional observations:

$$\mathcal{D} = \{(2,2),\ (1,1),\ (-1,-1),\ (-2,-2),\ (1,-1),\ (-1,1)\}.$$

The first four lie on the line $y = x$; the last two lie on $y = -x$. Consequently the cloud is *almost one-dimensional.* You may solve formally by computing the sample covariance matrix and its eigen-decomposition, *or intuitively by recognising the principal-component directions.*

1. **Centering.** Compute the sample mean $\boldsymbol{\mu}$ and the centred data matrix X_c.

2. **Covariance matrix.** Evaluate

$$\Sigma \ = \ \tfrac{1}{n-1}\, X_c^\top\, X_c.$$

3. **Eigen-decomposition.**
   (a) Find the eigenvalues $\lambda_1 \geq \lambda_2$ and *unit* eigenvectors $\mathbf{u}_1, \mathbf{u}_2$.
   (b) Identify which eigenvector is the *first* principal component (PC 1).

4. **Variance explained.**
   (a) What fraction of the total variance does PC 1 capture?
   (b) How many PCs are needed to retain at least 80% of the variance?

5. **Scores.** Project every centred observation onto PC 1 to obtain its 1-D scores.

6. **Low-rank reconstruction.** Using only PC 1, reconstruct each point

$$\hat{\mathbf{x}}_i \ = \ \boldsymbol{\mu} \ + \ \big(\mathbf{u}_1^\top(\mathbf{x}_i - \boldsymbol{\mu})\big)\,\mathbf{u}_1,$$

compute the mean-squared reconstruction error (MSRE), and relate it to the 20% variance that PC 1 fails to capture.

Hint: Because the data are already visually symmetric, much of the computation can be bypassed by reasoning about the principal directions $y = \pm x$.

4 Performance metrics

Consider a molecular test designed to determine whether a biological cell belongs to a specific cluster of malignant cells. Suppose the sensitivity and specificity of the test are both 95%, meaning that both false positives (the test indicates the cell is malignant when it is not) and false negatives (the test indicates the cell is not malignant when it actually is) occur in 5% of the cases. Despite this apparent precision, interpreting the test results still requires caution. Let's understand why:

(a) Someone claims that if a cell tests positive, then there is a 95% chance that it is indeed malignant. Is this statement correct? Explain.

(b) Now, consider that only 5% of the cells in a given tissue sample are actually malignant. What is the probability that a cell is malignant given that it tested positive? Interpret the result.
What happens if two independent tests are performed on the same cell and both return positive results (the test are also conditionally independent on the cancerous status of the cell)? Interpret the result.

(c) Suppose instead that 50% of the cells in the sample are malignant. How does this change the probability you computed in part (b)? Conclude what's the effect of prevalence, i.e. the fraction of cells that are malignant.

(d) We have studied the metrics you have found in (b) and (c). What's its name? Based on the information provided, can you compute the accuracy of the test? Justify your answer.

5 MCQs

5.1

MAP estimation maximizes $\log P(y|\beta) + \log P(\beta)$. For a Gaussian prior, this adds $-\frac{1}{2\tau^2}\sum_j \beta_j^2$ to the log-likelihood, thereby recovering:

(A) ridge regression's objective

(B) lasso regression's objective

(C) ordinary least squares

(D) elastic net with 50% mixing

5.2

Why does K-fold CV provide a more reliable error estimate than a single train/test split?

(A) it reduces bias but increases variance

(B) it uses each point once for validation, averaging over splits

(C) it always underestimates true error

(D) it replaces the need for regularization

5.3

In a GWAS with highly correlated SNPs, lasso selects one variant per LD block somewhat arbitrarily. This happens because:

(A) L1 regularization cannot handle any correlations

(B) the diamond-shaped L1 region intersects the RSS contours at a single corner

(C) lasso enforces all correlated features to zero simultaneously

(D) GWAS data violate the Laplace prior assumption

5.4

In a classification model with highly correlated gene-expression predictors, why does ridge regression often outperform OLS?

(A) it increases coefficient magnitudes for correlated predictors

(B) it penalizes large coefficients, stabilizing estimates under multicollinearity

(C) it drops one of each correlated pair automatically

(D) it guarantees unbiased estimates