BIOENG-210: Biological Data Science I: Statistical Learning

Theoretical Exercise Week 11
Prof. Gioele La Manno

May 2024

1 Uniqueness of the solution to the linear regression problem

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:

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

Solution:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) =& \nabla_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 = \\ &= (\nabla_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))(2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \\ &= 2\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})\end{aligned}$$

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 $\mathbf{A}$, $(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T\mathbf{A}$

Solution:

$$\nabla_{\boldsymbol{\beta}} (2X^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})) = \nabla_{\boldsymbol{\beta}} (2\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{X}$$

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).

Solution: We apply the condition for positive definiteness to the Hessian we computed in the previous step:

$$\mathbf{v}^T \mathbf{H} \mathbf{v} = \mathbf{v}^T (2\mathbf{X}^T \mathbf{X}) \mathbf{v} = 2(\mathbf{X}\mathbf{v})^T (\mathbf{X}\mathbf{v}) = 2||\mathbf{X}\mathbf{v}||_2^2 \geq 0$$

The only case in which this is equal to 0 is when $\mathbf{X}\mathbf{v} = \mathbf{0}$, which means that $\mathbf{v}$ is in the null space of $\mathbf{X}$, this can never happen if $\mathbf{X}$ 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 $\sigma(x) > 0, \forall x$.

Solution:

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \left(y_i \frac{d}{d\boldsymbol{\beta}} \log(\sigma(\mathbf{x}_i^T \boldsymbol{\beta})) + (1 - y_i) \frac{d}{d\boldsymbol{\beta}} \log(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})) \right) = \\
&= -\sum_{i=1}^{n} \left(y_i \frac{1}{\sigma(\mathbf{x}_i^T \boldsymbol{\beta})} \sigma(\mathbf{x}_i^T \boldsymbol{\beta})(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})) \mathbf{x}_i + (1 - y_i) \frac{-1}{1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})} \sigma(\mathbf{x}_i^T \boldsymbol{\beta})(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})) \mathbf{x}_i \right) = \\
&= -\sum_{i=1}^{n} \left(y_i (1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})) \mathbf{x}_i - (1 - y_i) \sigma(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i \right) = \sum_{i=1}^{n} (\sigma(\mathbf{x}_i^T \boldsymbol{\beta}) - y_i) \mathbf{x}_i
\end{aligned}$$

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

Solution:

$$\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big(\nabla_{\boldsymbol{\beta}}(\sigma(\mathbf{x}_i^T\boldsymbol{\beta}) - y_i)\mathbf{x}_i\big) = \\ = \sum_{i=1}^{n} \sigma(\mathbf{x}_i^T\boldsymbol{\beta})(1-\sigma(\mathbf{x}_i^T\boldsymbol{\beta}))\mathbf{x}_i\mathbf{x}_i^T$$

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 $\sigma(x)(1-\sigma(x)) \geq 0, \forall x$.

Solution: We will show that the Hessian is P.D by applying the condition shown above:

$$\mathbf{v}^T\mathbf{H}\mathbf{v} = \mathbf{v}^T\big(\sum_{i=1}^{n} \sigma(\mathbf{x}_i^T\boldsymbol{\beta})(1-\sigma(\mathbf{x}_i^T\boldsymbol{\beta}))\mathbf{x}_i\mathbf{x}_i^T\big)\mathbf{v} = \\ = \sum_{i=1}^{n} \sigma(\mathbf{x}_i^T\boldsymbol{\beta})(1-\sigma(\mathbf{x}_i^T\boldsymbol{\beta}))\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v} = \\ = \sum_{i=1}^{n} \sigma(\mathbf{x}_i^T\boldsymbol{\beta})(1-\sigma(\mathbf{x}_i^T\boldsymbol{\beta}))(\mathbf{v}^T\mathbf{x}_i)^2 \geq 0$$

In the last step we have combined the hint and the fact that $\mathbf{v}^T\mathbf{x}_i$ is a scalar, therefore $(\mathbf{v}^T\mathbf{x}_i)^2 \geq 0$.

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

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.

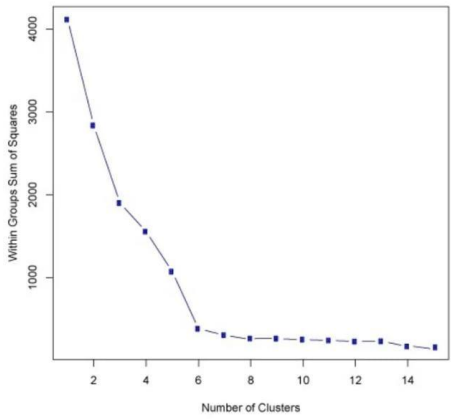

Figure 1: WCSS vs. number of clusters

Solutions:

(a) Two limitations of K-means clustering in high-dimensional biological data:

- **Assumption of spherical clusters:** K-means assumes that clusters are convex and isotropic (i.e., spherical in shape), which may not hold true for real biological data.
- **Sensitivity to initialization:** The algorithm can converge to different local minima depending on the initial positions of centroids. Multiple runs with different initializations are needed.
- **Distance metric limitations:** Euclidean distance becomes less meaningful in high-dimensional space due to the curse of dimensionality.

(b) Within-cluster sum of squares (WCSS):

$$J_{\text{WCSS}} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where μ_i is the centroid of cluster C_i. In the update step of K-means, the centroid is updated as the mean of all points in its cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

This choice of centroid minimizes the sum of squared Euclidean distances from each point to its cluster center.

(c) Elbow plot interpretation:

- As k increases, the WCSS decreases because each cluster has fewer points, and centroids can be placed more precisely.
- However, the rate of decrease slows after a certain point — adding more clusters yields only marginal improvement.

The **elbow point** is the value of k at which the decrease in WCSS begins to level off. It represents the optimal trade-off between:

- Minimizing intra-cluster variance (low WCSS), and
- Avoiding overfitting or excessive fragmentation (too many clusters).

In the figure, the elbow point appears to be at $k = 6$, where the WCSS starts to decrease at a slower rate. This suggests that six clusters are sufficient to capture the underlying structure of the data without overfitting.

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?

Solutions:
Cluster centroids:

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

Cell profiles:

$$x_A = [1.0,\ -0.6,\ 0.2]$$
$$x_B = [3.2,\ 1.4,\ -0.6]$$
$$x_C = [-1.0,\ -1.8,\ 1.3]$$

(a) Euclidean distances:

Cell A:

$$\begin{aligned} d_A(\mu_1) &= \sqrt{(1.0-1.2)^2 + (-0.6+0.5)^2 + (0.2-0.3)^2} \\ &= \sqrt{0.04 + 0.01 + 0.01} = \sqrt{0.06} \approx 0.245 \\ d_A(\mu_2) &= \sqrt{(1.0-3.5)^2 + (-0.6-1.1)^2 + (0.2+0.8)^2} \\ &= \sqrt{6.25 + 2.89 + 1.00} = \sqrt{10.14} \approx 3.183 \\ d_A(\mu_3) &= \sqrt{(1.0+0.9)^2 + (-0.6+2.0)^2 + (0.2-1.5)^2} \\ &= \sqrt{3.61 + 1.96 + 1.69} = \sqrt{7.26} \approx 2.695 \end{aligned}$$

Assign A to Cluster 1.

Cell B:

$$\begin{aligned} d_B(\mu_1) &= \sqrt{(3.2-1.2)^2 + (1.4+0.5)^2 + (-0.6-0.3)^2} \\ &= \sqrt{4.0 + 3.61 + 0.81} = \sqrt{8.42} \approx 2.902 \\ d_B(\mu_2) &= \sqrt{(3.2-3.5)^2 + (1.4-1.1)^2 + (-0.6+0.8)^2} \\ &= \sqrt{0.09 + 0.09 + 0.04} = \sqrt{0.22} \approx 0.469 \\ d_B(\mu_3) &= \sqrt{(3.2+0.9)^2 + (1.4+2.0)^2 + (-0.6-1.5)^2} \\ &= \sqrt{16.81 + 11.56 + 4.41} = \sqrt{32.78} \approx 5.725 \end{aligned}$$

Assign B to Cluster 2.

Cell C:

$$\begin{aligned} d_C(\mu_1) &= \sqrt{(-1.0-1.2)^2+(-1.8+0.5)^2+(1.3-0.3)^2} \\ &= \sqrt{4.84+1.69+1.00} = \sqrt{7.53} \approx 2.745 \\ d_C(\mu_2) &= \sqrt{(-1.0-3.5)^2+(-1.8-1.1)^2+(1.3+0.8)^2} \\ &= \sqrt{20.25+8.41+4.41} = \sqrt{33.07} \approx 5.75 \\ d_C(\mu_3) &= \sqrt{(-1.0+0.9)^2+(-1.8+2.0)^2+(1.3-1.5)^2} \\ &= \sqrt{0.01+0.04+0.04} = \sqrt{0.09} \approx 0.300 \end{aligned}$$

Assign C to Cluster 3.

(b) Silhouette coefficient for Cell A:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} = \frac{2.05-0.45}{\max(0.45,\ 2.05)} = \frac{1.60}{2.05} \approx 0.780$$

(c) Interpretation of silhouette drop (0.71 to 0.55): A drop in silhouette score indicates that the clusters have become less well-defined. This could mean that increasing k from 3 to 4 led to:

- Splitting of coherent clusters into smaller, less meaningful subgroups.
- Overfitting noise in the data.

Alternative method: Use the **elbow method** or inspect the **silhouette score across multiple** k values to choose the optimal number of clusters.

3 An almost-1-D cloud

Data set. Consider the six two-dimensional observations:

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

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 $\mathbf{u}_1, \mathbf{u}_2$ of the covariance matrix Σ.

(b) Identify which eigenvector is the *first* principal component (PC1).

4. **Variance explained.**

   (a) What fraction of the total variance does PC1 capture?

   (b) How many PCs are needed to retain at least 80 of the variance?

5. **Scores.** Project every centred observation onto PC1 to obtain its 1-D scores.

6. **Low-rank reconstruction.** Using only PC1, reconstruct each point

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

   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 $y = \pm x$.

Solutions:

We annotate each step; an intuitive route that spots the PCs first yields the same answers more quickly.

1. Centering The data are symmetric about the origin, hence

$$\boldsymbol{\mu} = (0,0)^\top, \qquad X_c = X.$$

2. Covariance matrix Forming $X_c^\top X_c$ and dividing by $n-1 = 5$ gives

$$\Sigma \;=\; \begin{pmatrix} 2.4 & 1.6 \\ 1.6 & 2.4 \end{pmatrix}.$$

Intuition. The large positive off-diagonal element already signals that the major axis lies along $y = x$.

3. Eigen-decomposition Solving $\det(\Sigma - \lambda I) = 0$:

$$(2.4 - \lambda)^2 - 1.6^2 = 0 \to \lambda = 2.4 \pm 1.6$$

$$\lambda_1 = 4.0, \quad \lambda_2 = 0.8.$$

Corresponding unit eigenvectors:

$$\mathbf{u}_1 = \tfrac{1}{\sqrt{2}}(1,\,1)^\top, \qquad \mathbf{u}_2 = \tfrac{1}{\sqrt{2}}(1,\,-1)^\top.$$

So PC 1 runs along the 45° line $y = x$ and PC 2 is the perpendicular axis.

4. Variance explained Total variance = $\mathrm{tr}(\Sigma) = 4.8$.

$$\frac{\lambda_1}{\lambda_1+\lambda_2} \;=\; \frac{4}{4.8} \;=\; 0.833 \;\; (83.3\%).$$

Hence one component already exceeds the 80 target.

5. Scores For each centred point (x_i, y_i),

$$s_i = \mathbf{u}_1^\top (x_i, y_i)^\top = \tfrac{1}{\sqrt{2}}(x_i + y_i).$$

This yields $s = \{+2\sqrt{2},\, +\sqrt{2},\, -\sqrt{2},\, -2\sqrt{2},\, 0,\, 0\}$ for the six observations.

6. Low-rank reconstruction and MSRE With only PC 1,

$$\hat{\mathbf{x}}_i = s_i\, \mathbf{u}_1.$$

The first four points (on $y = x$) are recovered perfectly; the two on $y = -x$ collapse to $(0, 0)$, each incurring squared error 2.

$$\mathrm{MSRE} = \frac{0+0+0+0+2+2}{6} = \frac{4}{6} \approx 0.66,$$

Intuitive view. Visually, the cloud is a "cigar" whose long axis is $y = x$; projecting onto that line keeps four points unchanged and folds the other two to the centre, exactly as the numerical PCA confirms.

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.

(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.

Solution

We provide the solution in terms of the probability, although one could use alternate approaches.
Reminder: the prevalence is the fraction of positive in the population.

Define M ('malignant') as the event that a cell truly belongs to the malignant cluster, and P ('positive') as the event that the test result is positive. The sensitivity and specificity of the test mean that

$$P(P \mid M) = P(N \mid m) = 0.95.$$

(a) The claim to be evaluated is whether $P(M \mid P) = 0.95$. In the problem statement, we are only told that $P(P \mid M) = 0.95$, so there is no reason to believe it.

(b) Suppose we are told that only 5% of the cells in a sample are malignant. Then

$$P(M) = \frac{5}{100}, \quad P(m) = \frac{95}{100}.$$

By Bayes' theorem

$$P(M \mid P) = \frac{P(P \mid M) \cdot P(M)}{P(P)}.$$

We find $P(P)$ using the law of total probability

$$P(P) = P(P \mid M) \cdot P(M) + P(P \mid m) \cdot P(m) = 0.95 \cdot 0.05 + 0.05 \cdot 0.95 = 0.0475 + 0.0475 = 0.095.$$

Then

$$P(M \mid P) = \frac{0.95 \cdot 0.05}{0.095} = \frac{1}{2}.$$

Thus, even though the test is 95% accurate in both directions, a single positive result only gives a 50% chance of malignancy due to the low base rate.

We now consider taking *two independent tests* and both return positive.

$$P(M \mid P_1 \cap P_2) = \frac{P(P_1 \cap P_2 \mid M) \cdot P(M)}{P(P_1 \cap P_2)}.$$

Where

$$P(P_1 \cap P_2 \mid M) = P(P_1 \mid M)^2 = 0.95^2, \quad P(P_1 \cap P_2 \mid m) = 0.05^2.$$

Then

$$P(P_1 \cap P_2) = P(P_1 \cap P_2 \mid M) \cdot P(M) + P(P_1 \cap P_2 \mid m) \cdot P(m)$$

Finally

$$P(M \mid P_1 \cap P_2) \approx 0.95.$$

So, two positive tests bring the posterior probability of malignancy back to 95%, illustrating how repeated independent tests can significantly increase certainty.

(c) If instead $P(M) = 0.5$, we recompute:

$$P(P) = 0.95 \cdot 0.5 + 0.05 \cdot 0.5 = 0.5,$$

$$P(M \mid P) = \frac{0.95 \cdot 0.5}{0.5} = 0.95.$$

Higher prevalence leads to higher confidence in a positive result.

(d) What we computed is nothing but

$$\text{Precision} = P(\text{M} \mid \text{P}) \hat{=} \frac{TP}{TP + FP}$$

From the initial description alone, we cannot compute precision and accuracy. However, once we are given the prevalence, it is possible to obtain such values as you have seen for the precision and as follows for the accuracy

$$\text{Accuracy} = \text{Sensitivity} \cdot \text{Prevalence} + \text{Specificity} \cdot (1 - \text{Prevalence}).$$

MCQs

- **MAP estimation maximizes**

$$\log P(\mathbf{y} \mid \boldsymbol{\beta}) + \log P(\boldsymbol{\beta}).$$

For a Gaussian prior this adds a penalty term

$$-\frac{1}{2\tau^2} \sum_j \beta_j^2$$

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

Solution:
A. A Gaussian prior on each coefficient, $\beta_j \sim \mathcal{N}(0, \tau^2)$, contributes $\log P(\beta_j) \propto -\frac{1}{2\tau^2}\beta_j^2$. Summing over j gives an L_2 penalty, so maximizing the posterior is equivalent to minimizing least-squares plus an $\alpha\|\beta\|_2^2$ term—that is, ridge regression.

- **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

  Solution:
  B. With a single split, the test error can swing widely depending on which samples end up in the test set. K-fold CV cycles each subset through being "held out" exactly once, then averages the K validation errors—yielding a lower-variance, more robust estimate.

- **In a GWAS with highly correlated SNPs, lasso often selects only one variant per LD block somewhat arbitrarily. Why?**

  A. L1 regularization cannot handle any correlations

  B. the diamond-shaped L_1 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

  Solution:
  B. Geometrically, the constraint $\|\beta\|_1 \leq t$ is a "diamond" in coefficient space and the RSS contours are ellipses. Their first point of contact is almost always at a corner—i.e. one nonzero β—so lasso picks a single SNP in each correlated block.

- **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

  Solution:
  B. OLS estimates blow up when predictors are collinear (high variance). Adding an L_2 penalty shrinks coefficients toward zero, taming variance at the cost of a little bias—and thus improving overall generalization under multicollinearity.