COMP3670/6670: Introduction to Machine Learning

Release Date. October 5, 2023
Due Date. 11:59 pm, October 23, 2023
Maximum credit. 100

Question 1 **Properties of Eigenvalues** ($10 + 5 = 15$ credits)

Let $\mathbf{A}$ be an invertible matrix.

1. (a) Prove that all the eigenvalues of $\mathbf{A}$ are non-zero.
   (b) Prove that for any eigenvalue λ of $\mathbf{A}$, λ^{-1} is an eigenvalue of $\mathbf{A}^{-1}$.
2. Let $\mathbf{B}$ be a square matrix. Let $\mathbf{x}$ be an eigenvector of $\mathbf{B}$ with eigenvalue λ. Prove that for all integers $n \geq 1$, $\mathbf{x}$ is an eigenvector of $\mathbf{B}^n$ with eigenvalue λ^n.

Question 2 **Distinct eigenvalues and linear independence** ($10 + 5 = 15$ credits)

Let $\mathbf{A}$ be a $n \times n$ matrix.

1. Suppose that $\mathbf{A}$ has n distinct eigenvalues $\lambda_1, \dots, \lambda_n$, and corresponding non-zero eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$. Prove that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are linearly independent.

   Hint: You may use without proof the following property: If $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ are linearly dependent then there exists some p such that $1 \leq p < m$, $\mathbf{y}_{p+1} \in \text{span}\{\mathbf{y}_1, \dots, \mathbf{y}_p\}$ and $\{\mathbf{y}_1, \dots, \mathbf{y}_p\}$ are linearly independent.
2. Hence, or otherwise, prove that for any matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, there can be at most n distinct eigenvalues for $\mathbf{B}$.

Question 3 **Determinants** ($5 + 5 + 5 + 5 + 5 = 25$ credits)

1. Compute the determinant of matrix $\mathbf{A}$ using the Sarus rule

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & -2 \\ 0 & -1 & -2 \end{bmatrix}$$

2. Prove $\mathbf{det}(\mathbf{A}^T) = \mathbf{det}(\mathbf{A})$.

3. Prove $\mathbf{det}(\mathbf{I}_n) = 1$ where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

4. Prove $\mathbf{det}(\mathbf{A}) = -\mathbf{det}(\sigma_{i,j}(\mathbf{A}))$ for $i \neq j$ where $\sigma_{i,j}$ swaps the i'th and j'th row of the input matrix.

5. Let $\mathbf{U}$ be an square $n \times n$ **upper** triangular matrix. Prove that the determinant of $\mathbf{U}$ is equal to the product of the diagonal elements of $\mathbf{U}$.

Question 4 **Trace Inequality** (10 credits)

Prove for arbitrary square matrix $\mathbf{A}$ and $\mathbf{B}$,

$$\mathbf{tr}\left((\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})^T\right) \leq 2 \cdot \mathbf{tr}(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)$$

Question 5 **Computations with Eigenvalues** ($3 + 3 + 3 + 3 + 3 = 15$ credits)

Let $\mathbf{A} = \begin{bmatrix} 2 & -2 \\ 0 & 1 \end{bmatrix}$.

1. Compute the eigenvalues of $\mathbf{A}$.

2. Find the eigenspace E_λ for each eigenvalue λ. Write your answer as the span of a collection of vectors.

3. Verify the set of all eigenvectors of $\mathbf{A}$ spans $\mathbb{R}^2$.

4. Hence, find an invertable matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$.

5. Hence, find a formula for efficiently [1] calculating $\mathbf{A}^n$ for any integer $n \geq 0$. Make your formula as simple as possible.

Question 6 **PCA as an optimisation problem** (20 credits)

Principal component analysis (PCA) is a technique for increasing the interpretability of data of datasets with a large number of dimensions/features per observation while preserving the maximum amount of information. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data. [2]

A common technique used for finding a local max/min of a function $f(x)$ subject to an equality constraint of the form $g(x) = c$, is using a Lagrange multiplier. Naïvely, to find a local min or max we would start by differentiating $f(x)$ to get $\frac{df}{dx}$ and then solve for

$$\frac{df}{dx} = 0$$

[1] That is, a closed form formula for $\mathbf{A}^n$ as opposed to multiplying $\mathbf{A}$ by itself n times over.

[2] `https://en.wikipedia.org/wiki/Principal_component_analysis`

But this does not take into account the constraint we have imposed. In order to account for this, we instead create the expression, called the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) - \lambda(g(x) - c)$$

Then to get the local min/max, we differentiate this expression with respect to both x and λ and set both equal to zero. So we have the equations,

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 0$$

Solving the simultaneous equation given by setting these two expressions equal to zero, gives the local minimum/maximum of $f(x)$, subject to the constraint $g(x) = c$. (You can read more into Lagrange multipliers if you want but it is not needed for the question)

In this question, our goal is to motivate why the vectors chosen for PCA (the principal components) are eigenvectors of the covariance matrix. Suppose we have a dataset $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ of vectors with mean centered at 0 (for simplicity, we can always adjust by subtracting the mean if needed). For a given vector v, the result of projecting a vector x_i onto v is given by $v^\intercal x_i$. The variance of the projected data is given by $\frac{1}{n}\sum_{i=1}^{n}(v^\intercal x_i - \mu)^2$ where μ is the mean of the projected data. Since we are assuming the mean of the data is 0, this simplifies to $\frac{1}{n}\sum_{i=1}^{n}(v^\intercal x_i)^2$. Our goal is to find the vector v that maximises this variance.

$$\begin{aligned}
\mathcal{V} = \frac{1}{n}\sum_{i=1}^{n}(v^\intercal x_i)^2 &= \frac{1}{n}\sum_{i=1}^{n} v^\intercal x_i x_i^\intercal v \\
&= v^\intercal \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\intercal\right) v \\
&= v^\intercal \mathcal{C} v
\end{aligned}$$

Use the Lagrangian method to show that the vector v that maximises the resulting variance $\mathcal{V} = v^\intercal \mathcal{C} v$ subject to the constraint $\|v\|_2 = 1$ is an eigenvector of the covariance matrix $\mathcal{C}$.