The Australian National University Semester 2, 2023
School of Computing Assignment 4 Theory Questions

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.

   Solution. If $\lambda = 0$ is an eigenvalue of $\mathbf{A}$, then the equation $\mathbf{Ax} = \mathbf{0}$ has non-trivial solutions. But, multiplying by $\mathbf{A}^{-1}$, we obtain

   $$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{0}$$

   $$\mathbf{x} = \mathbf{0}$$

   a contradiction.

   (b) Prove that for any eigenvalue λ of $\mathbf{A}$, λ^{-1} is an eigenvalue of $\mathbf{A}^{-1}$.

   Solution. Let λ be an eigenvalue of $\mathbf{A}$. Then there exists $\mathbf{x} \neq \mathbf{0}$ such that

   $$\mathbf{Ax} = \lambda\mathbf{x}$$

   $$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\lambda\mathbf{x}$$

   $$\mathbf{x} = \lambda\mathbf{A}^{-1}\mathbf{x}$$

   $$\frac{1}{\lambda}x = \mathbf{A}^{-1}\mathbf{x}$$

   $$\mathbf{A}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$$

   where the second to last step is justified by 1a, as $\lambda \neq 0$.

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.

Solution. We proceed by induction on n.
Base case, $n = 1$ is trivial, as we are already given that λ is an eigenvalue of $\mathbf{B}$. Step case. Assume that λ^n is an eigenvalue of $\mathbf{B}^n$. Let $\mathbf{x}$ be the corresponding eigenvector for λ^n. Then,

$$\mathbf{B}^{n+1}\mathbf{x} = \mathbf{B}\mathbf{B}^n\mathbf{x} = \mathbf{B}(\mathbf{B}^n x) = \mathbf{B}\lambda^n x = \lambda^n(\mathbf{B}x) = \lambda^n\lambda\mathbf{x} = \lambda^{n+1}x$$

and hence λ^n is an eigenvalue of $\mathbf{B}^n$ for all $n \geq 1$.

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.

   Solution. Suppose for a contradiction that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is linearly dependent. Then, by the hint, there exists p such that $1 \leq p < n$, $\mathbf{x}_{p+1} \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ and $\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ is linearly independent. So,

   $$\mathbf{x}_{p+1} = c_1\mathbf{x}_1 + \ldots + c_p\mathbf{x}_p \tag{1}$$

   for some collection of scalars $c_1, \dots, c_p$. Apply A to both sides of the above equation, (noting that $\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$) and obtain

   $$\lambda_{p+1}\mathbf{x}_{p+1} = c_1\lambda_1\mathbf{x}_1 + \ldots + c_p\lambda_p\mathbf{x}_p \tag{2}$$

   We can also multiply both sides of the first equation by λ_{p+1}, and obtain

   $$\lambda_{p+1}\mathbf{x}_{p+1} = c_1\lambda_{p+1}\mathbf{x}_1 + \ldots + c_p\lambda_{p+1}\mathbf{x}_p \tag{3}$$

   Equations (2) and (3) are equal to each other.

   $$c_1\lambda_1\mathbf{x}_1 + \ldots + c_p\lambda_p\mathbf{x}_p = c_1\lambda_{p+1}\mathbf{x}_1 + \ldots + c_p\lambda_{p+1}\mathbf{x}_p$$
   $$c_1(\lambda_1 - \lambda_{p+1})\mathbf{x}_1 + \ldots + c_p(\lambda_1 - \lambda_{p+1})\mathbf{x}_p = 0$$

   Since each of the eigenvalues are distinct, $\lambda_i - \lambda_{p+1} \neq 0$ for all $1 \leq i \leq p$. Since $\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ are linearly independent, the only way that $c_1(\lambda_1 - \lambda_{p+1})\mathbf{x}_1 + \ldots + c_p(\lambda_1 - \lambda_{p+1})\mathbf{x}_p = 0$ could be true is if $c_1 = \ldots = c_p = 0$. Therefore, by Equation (1), $\mathbf{x}_{p+1} = 0$, a contradiction.

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 $\mathbf{B}$.

   Solution. Suppose not. Then $\mathbf{B}$ has m distinct eigenvalues $\{\lambda_1, \dots, \lambda_m\}$ with $m > n$. Let $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ denote the corresponding eigenvectors. Then $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ is linearly independent by the previous theorem. But that would mean we have m many linearly independent vectors in $\mathbb{R}^n$ with $m > n$, a contradiction.

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

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

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

Solution. According to Sarrus rule

$$\mathbf{det}(\mathbf{A}) = 1 \cdot 3 \cdot (-2) + 2 \cdot (-1) \cdot 4 + 0 - 0 - 1 \cdot (-1) \cdot (-2) - 2 \cdot 2 \cdot (-2) = -8$$

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

Solution. For a given square matrix A of larger dimension, the **minor** M_{ij} of entry a_{ij} is the determinant of the submatrix obtained by deleting the i^{th} row and the j^{th} column from A. Let $C_{ij} := (-1)^{i+j} M_{ij}$ denote the cofactor of a_{ij}. By choosing any row i or column j of A, the determinant is defined recursively, by cofactor expansion along that row or column.

$$\det A = \sum_k a_{kj} C_{kj} = \sum_k a_{ik} C_{ik}$$

By the definition of the determinant, we get the same answer whether we cofactor expand along a row or a column, and cofactor expanding along the first row of A is equivalent to cofactor expanding along the first column of A^T, so it must be the case that

$$\det(A) = \det(A^T)$$

Note: If the student assume the matrix $\mathbf{A}$ is the given one in the previous question, it's fine.

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

Solution. Proof by induction. Trivial for $n = 1$. Assume true for some n. Then by cofactor expanding along the first column of I_{n+1},

$$\det I_{n+1} = \sum_{k=1}^{n+1} I_{1k} C_{1k} = I_{11} C_{11} + \sum_{k=2}^{n+1} I_{1k} C_{1k}$$

All off diagonals terms of I_{n+1} are zero, all diagonal terms are one. Therefore,

$$\det I_{n+1} = 1 \times C_{11} = (-1)^{1+1} M_{11} = \det I_n = 1$$

Hence result follows by induction.

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.

Solution. By definition,

$$\sigma_{i,j}(\mathbf{A}) = \sigma_{i,j}(\mathbf{I}) \cdot \mathbf{A}$$

as row swapping is an elementary row operation. Hence,

$$\mathbf{det}(\sigma_{i,j}(\mathbf{A})) = \mathbf{det}(\sigma_{i,j}(\mathbf{I})) \cdot \mathbf{det}(\mathbf{A})$$

Now the goal becomes to prove

$$\mathbf{det}(\sigma_{i,j}(\mathbf{I})) = -1$$

This can be proved easily by induction.

Note: If the student assume the matrix $\mathbf{A}$ is the given one in the previous question, it's fine.

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 $\mathbf{U}$.

Solution. Proceed by induction on n. Trivial for $n = 1$. Assume true for all square upper triangular matrices U of dimension $n \times n$. For an arbitrary $(n+1) \times (n+1)$ upper triangular matrix V, cofactor expand along the first column. As V is upper triangular, all terms in the first column apart from U_{11} are zero. Thus,

$$\det V = \sum_k V_{k1} C_{k1}$$

where C_{ij} is the cofactor along the i^{th} row and j^{th} column of V. Thus,

$$\det V = \sum_k V_{k1} C_{k1} = V_{11} C_{11} + \sum_{k \neq 1} V_{k1} C_{k1}$$

$$= V_{11} C_{11} = V_{11} (-1)^{1+1} M_{11} = V_{11} \det U$$

where U is the upper triangular matrix formed by deleting the first row and first column of V. Since the dimension of U is $n \times n$, apply induction hypothesis to obtain

$$\det U = V_{22} \times V_{33} \times \ldots \times V_{(n+1)(n+1)}$$

and therefore

$$\det V = V_{11} \times V_{22} \times \ldots \times V_{(n+1)(n+1)} = \prod_{i=1}^{n+1} V_{ii}$$

as required.

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

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

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

Solution. Denote row i of $\mathbf{A}$, $\mathbf{B}$ as $\mathbf{a}_i^T$, $\mathbf{b}_i^T$.

$$\begin{aligned} LHS - RHS &= \mathbf{tr}(\mathbf{AA}^T + \mathbf{AB}^T + \mathbf{BA}^T + \mathbf{BB}^T) - 2 \cdot \mathbf{tr}(\mathbf{AA}^T + \mathbf{BB}^T) \\ &= \mathbf{tr}(\mathbf{AB}^T + \mathbf{BA}^T - \mathbf{AA}^T - \mathbf{BB}^T) \\ &= \sum_{i=1}^{n} \mathbf{a}_i^T \mathbf{b}_i + \mathbf{b}_i^T \mathbf{a}_i - \mathbf{a}_i^T \mathbf{a}_i - \mathbf{b}_i^T \mathbf{b}_i \\ &= -\sum_{i=1}^{n} (\mathbf{a}_i - \mathbf{b}_i)^T (\mathbf{a}_i - \mathbf{b}_i) \leq 0 \end{aligned}$$

Hence,

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

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 $\mathbf{A}$.

   Solution.

   $$\mathbf{det}(\mathbf{A} - \lambda \mathbf{I}) = (2-\lambda)(1-\lambda) = 0$$

   Thus, $\lambda_1 = 1$, $\lambda_2 = 2$.

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

   Solution. For $\lambda_1 = 1$:

   $$\begin{bmatrix} 2 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

   The solution space is

   $$E_1 = \left\{ x_2 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}, x_2 \in \mathbb{R} \right\} = \mathbf{Span} \left\{ \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\}$$

   For $\lambda_2 = 2$:

   $$\begin{bmatrix} 2 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

   The solution space is

   $$E_2 = \left\{ x_1 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_1 \in \mathbb{R} \right\} = \mathbf{Span} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

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 $\mathbf{A} = \mathbf{PDP}^{-1}$.

Solution. Consider the matrix corresponding to the span

$$\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$$

This is invertible as its determinant is $-1 \neq 0$. Hence eigenvectors span $\mathbb{R}^2$.

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.

   Solution. From the week 8? tutorials, we proved that if $\mathbf{A} = \mathbf{PDP}^{-1}$, then $\mathbf{A}^n = \mathbf{PD}^n\mathbf{P}^{-1}$. Since $\mathbf{D}$ is a diagonal matrix, we have

$$\mathbf{D}^n = \begin{bmatrix} 2^n & 0 \\ 0 & 1 \end{bmatrix}$$

   Hence,

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2^n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 2^{n+1} - 2 \\ 0 & 2^n \end{bmatrix}$$

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

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

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,

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

[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

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 $\mathcal{C}$.

Solution. Phrase the problem as an optimisation problem:

$$\begin{aligned}\max_{\mathbf{v}} \quad & \mathbf{v}^T\mathbf{C}\mathbf{v} \\ \text{s.t.} \quad & \|\mathbf{v}\|_2 = 1\end{aligned}$$

The square root in the constraint is not so easy to differentiate. Hence, we consider an equivalent problem

$$\begin{aligned}\max_{\mathbf{v}} \quad & \mathbf{v}^T\mathbf{C}\mathbf{v} \\ \text{s.t.} \quad & \mathbf{v}^T\mathbf{v} = 1\end{aligned}$$

Note this is a non-convex optimisation problem. Its Lagrangian should be

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T\mathbf{C}\mathbf{v} - \lambda(\mathbf{v}^T\mathbf{v} - 1)$$

If $(\mathbf{v}^*, \lambda^*)$ is optimal, they should satisfy the first order condition

$$\nabla\mathcal{L}(\mathbf{v}^*, \lambda^*) = \mathbf{0} \qquad \mathbf{v}^{*T}\mathbf{v}^* = 1$$

First we set the gradients to 0, as $\mathbf{C}$ is a covariance, $\mathbf{C}$ is always symmetric:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{v}}|_{\mathbf{v}^*,\lambda^*} = \mathbf{v}^{*T}(\mathbf{C} + \mathbf{C}^T) - 2\lambda^*\mathbf{v}^{*T} = 2\mathbf{v}^{*T}\mathbf{C} - 2\lambda^*\mathbf{v}^{*T} = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda}|_{\mathbf{v}^*,\lambda^*} = 1 - \mathbf{v}^{*T}\mathbf{v}^* = 0$$

By simplifying the first equation, we have

$$\mathbf{C}\mathbf{v}^* = \lambda^*\mathbf{v}^*$$

This clearly shows that if $(\mathbf{v}^*, \lambda^*)$ are optimal, $\mathbf{v}^*$ must be an eigenvector of $\mathbf{C}$ with associated eigenvalue λ^* (KKT condition). In this case, the variance is

$$\mathbf{v}^{*T}\mathbf{C}\mathbf{v}^* = \lambda^*$$

The Lagrangian gives us a set of potential solutions: the eigenvectors. Among those eigenvectors, we choose the one that maximises the objective function. Clearly, we choose the eigenvector with the largest eigenvalue as our goal here is to maximise.