Machine Learning, Tutorial 1
University of Bern

Mehdi Noroozi (noroozi@inf.unibe.ch)

19/09/2018

Linear algebra

1. Consider the matrix $G = A^\top A$, where $A \in \mathbb{R}^{m\times n}$. Show that G is positive semi-definite.
Solution.
$$x^\top Gx = x^\top A^\top Ax = (Ax)^\top (Ax) = y^\top y \geq 0$$

2. Show that a positive definite matrix is non-singular.
Solution. Suppose A is positive definite and non-singular. So it is not full rank.Then there is a column of A, like j, that can be expressed as a linear combination of the other columns.
$$a_j = \sum_{i\neq j} x_i a_i$$
setting $x_j = -1$, we have
$$Ax = \sum_{i=1}^{n} x_i a_i = 0$$
But this implies that $x^\top Ax = 0$

3. Show that if (λ_i, x_i) are the i-th eigenvalue and i-th eigenvector of a non-singular and symmetric matrix $A \in \mathbb{R}^{n\times n}$, then $(\frac{1}{\lambda_i}, x_i)$ are the i-th eigenvalue and i-th eigenvector of A^{-1}.

Solution.
$$Ax = \lambda_i x_i \rightarrow x_i = \lambda_i A^{-1} x_i \rightarrow A^{-1} x_i = \frac{1}{\lambda_i} x$$

4. Given two sets of vectors $\{x_1, ...x_n\} \subset \mathbb{R}^n$ and $\{y_1, ..., y_n\} \subset \mathbb{R}^n$, show that rank $\left[\sum_{i=1}^{m} x_i y_i^\top\right] \leq m$.
Hint: First show that the square matrix $x_i y_i^\top$ has rank 1.

Solution.
$$A^{(i)} = x_i y_i^\top = \begin{bmatrix} -a_1^{(i)}- \\ \vdots \\ -a_n^{(i)}- \end{bmatrix}. \tag{1}$$
But $a_1^{(i)} = (x_i)_1 y_i, ..., a_m^{(i)} = (x_i)_n y_i$. So every column of $A^{(i)}$ can b obtained by scalar product of y_i. This implies that $A^{(i)}$ is of rank 1.
We know that $rank(A + B) \leq rank(a) + rank(B)$. Considering that $rank(x_i y_i^\top) = 1, \forall i$ we have rank $\left[\sum_{i=1}^{m} x_i y_i^\top\right] \leq m$.

5. Show that rank$(A) \leq \min\{m, n\}$, where $A \in \mathbb{R}^{m\times n}$.

Solution. We know that column rank and row rank of any matrix is equal. also we know that the column rank is at most equal to the number of columns and the row rank is at most equal to the number of rows. These two consideration implies that rank$(A) \leq \min\{m, n\}$.

6. In each of the following cases, state whether the matrix A is guaranteed to be non-singular or not. Justify your answer in each case.

   (a) $A \in \mathbb{R}^{m\times n}$ is a full rank matrix.
   Solution. No. A non-singular matrices should be square.

   (b) $|A| = 0$.
   Solution. No. An square matrix A is non-singulare if and only $|A| \neq 0$

   (c) A is an orthogonal matrix.
   Solution. Yes, For an orthogonal matrix Q we have $Q^\top Q = QQ^\top = I$ so Q is non singular and $Q^{-1} = Q^\top$

   (d) A has no eigenvalue equal to zero.
   Solution. Yes, we know that $|A| = \prod \lambda_i$. So if A has no zero eigenvalue, then $|A| \neq 0$ so A is non-singular.

   (e) A is a symmetric matrix with non-negative eigenvalues.
   Solution. No. we know that if A is a symmetric matrix then $x^\top A = \sum_{i=1}^{n} \lambda_i y_i^2$ and it is positive/negative for any x if and only if all the eigenvalues are positive/negative.

Calculus

The Gradient and Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ is calculated as the following:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1^2} & \frac{\partial f(x)}{\partial x_1 x_2} & \dots & \frac{\partial f(x)}{\partial x_1 x_n} \\ \frac{\partial f(x)}{\partial x_2 x_1} & \frac{\partial f(x)}{\partial x_2^2} & \dots & \frac{\partial f(x)}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x)}{\partial x_1 x_n} & \frac{\partial f(x)}{\partial x_2 x_n} & \dots & \frac{\partial f(x)}{\partial x_n^2} \end{bmatrix}$$

1. Calculate the Gradient and Hessian of the function $f : \mathbb{R}^n \to \mathbb{R}$ in the following cases where $a, b \in \mathbb{R}^n$ and Q is a $n \times n$ symetirc matrix

   - $f(x) = a^\top x$
   Solution. $\frac{\partial b^\top x}{\partial x_i} = b_i \to \nabla_x b^\top x = b$ The hessian is equal to 0 as the gradient is a constant vector.
   - $f(x) = \frac{1}{2} Q^\top x Q - b^\top x$

$$x^\top Q x = \sum_{i=1}^{n} x_i (Qx)_i) = \sum_{i=1}^{n} x_i (\sum_{j=1}^{n} Q_{ij} x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} x_i x_j.$$

$$\begin{aligned} \frac{\partial x^\top Q x}{\partial x_k} &= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j=1}^{n} Q_{ij} x_i x_j + \sum_{j=1}^{n} Q_{kj} x_k x_j\right] \\ &= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j\neq k} Q_{ij} x_i x_j + \sum_{i\neq k} Q_{ik} x_i x_k + \sum_{j\neq k} Q_{kj} x_k x_j + Q_{kk} x_k^2\right] \\ &= \sum_{i\neq k} Q_{ik} x_i + \sum_{j\neq k} Q_{kj} x_j + 2Q_{kk} x_k \\ &= \sum_{i=1}^{n} Q_{ik} x_i + \sum_{j=1}^{n} Q_{kj} x_j = 2\sum_{i=1}^{n} Q_{ki} x_i \end{aligned}$$

   So the $\frac{\partial x^\top Q x}{\partial x_k} = q_k^\top x$ where q_k is the q-th row vector of Q and $\nabla_x x^\top Q x = 2Qx$.
   Therefore, $\nabla_x[\frac{1}{2} x^\top Q x - b^\top x] = Qx - b$.
   Following the definiton of Hessian, one can easily show that it is equla to Q.

Probability

1. In this exercise we analyse the Monty Hall game. The rules of the game are the following. There are 3 doors A, B and C. Behind one of the doors there is the great prize, a sport-car. Behind other doors there is nothing. Firstly, the player chooses one of the doors. Secondly the anchorman (Monty Hall) opens one of the doors the player did not choose, such that the sport-car is not behind that door. The anchorman can always do this, because he knows where the sport-car is. At this point the player has to choose whether he sticks with her original choice, or changes to the other door. She will receive whatever behind the door she chooses. What is the best strategy to win the prize?

   Solution.
   There are two possibilities. The prize can be behind the door we first choose with probability $1/3$ or it can be behind the other two doors with probability $2/3$. When we stick to our first choice, there is $1/3$ probability that we get the prize. The better strategy is to switch to the other door and the probability of winning the prize is :

$$P(\text{"win with change"}) = P(\text{"win with change"} \cap \text{"the prize is in the current selection"}) + \\ P(\text{ "win with change"} \cap \text{"the prize is in not in the current selection"}) = \frac{1}{3} \times 0 + \frac{2}{3} \times 1 = \frac{2}{3}$$

2. Given the following statistics what is the probality that a woman has cencer if she has a positive mammogram resutl?
   - One percent of women over 50 have breast cancer.
   - Ninety prcent of women who have breast cancer test positive on mammmograms.
   - Eight precent of women will have false positives.

   Solution.

$$P(H|e) = \frac{P(H,e)}{P(e)} = \frac{P(e|H)P(H)}{P(e \cap H) + P(e \cap \sim H)} = \frac{P(e|H)P(H)}{P(e|H)P(H) + P(e| \sim H)P(\sim H)}$$
$$= \frac{0.9 \times .01}{.9 \times .01 + .08 \times 0.99} \approx 0.1$$

3. Thomas and Viktor are friends. It is a friday night and Thomas does not have phone. Viktor knows that there is $2/3$ probability that Thomas goes to party in downtown. There are 5 pubs in downtown and there is equal probability of Thomas going to them if he goes to party. Viktor already looked for Thomas in 4 of the bars. What is the probability of Viktor finding Thomas in the last bar?

   Solution.
   The sample space is

$$S = \{\text{"home"}, \text{"pub 1"}, \text{"pub 2"}, \text{"pub 3"}, \text{"pub 4"}, \text{"pub 5"}\},$$

   and the probability of the events are $P(\text{"home"}) = 1/3$ and $P(\text{"pub i"}) = 2/15$. We need to compute $P(\text{"pub 5"}|\text{"not in pub 1 ... 4"})$. Using the hayes rule,

$$P(\text{"pub 5"}|\text{"not in pub 1 ... 4"}) = \\ \frac{P(\text{"pub 5"} \cap \text{"not in pub 1 ... 4"})}{P(\text{"not in pub 1 ... 4"})} = \frac{2/15}{7/15} = \frac{2}{7}.$$

   Note that :

$$P(\text{"not in pub 1 ... 4"}) = P(\text{"home"} and \text{"not in pub 1 ... 4"}) + P(\text{"out"} and \text{"not in pub 1 ... 4"}) = \frac{1}{3} \times 1 + \frac{2}{3} \times \frac{1}{5} = \frac{7}{15}$$

4. Prove the following statements regarding the covariance:

   (a) Show that if X and Y are independent then $\text{Cov}[X, Y] = 0$.
   Give an example that shows that the opposite is not true.
   Solution

$$\text{Cov}[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY - XE[Y] - YE[X] + E[X]E[Y]]$$
$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] = E[XY] - E[X]E[Y]$$

   If X, Y are independent, $E[XY] = E[X]E[Y]$ therefore the covariance is equal to 0. To show that the opposite is not correct, consider random variales $X \sim \mathcal{N}(0,1), Y = X^2$ we know that $E(X) = E(X^3) = 0$. Threfore, $Cov(X,Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2] = 0$. Howiver, it is clear that X, Y are not independent.

(b) Show that the covariance matrix is always symmetric and positive semidefinite.

Solution

The $(i,j)^{th}$ element of the covariance matrix Σ is given by

$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X_j - \mu_j)(X_i - \mu_i)] = \Sigma_{ji}$$

so that the covariance matrix is symmetric.

For an arbitrary vector u,

$$\begin{aligned} u^\top \Sigma u &= u^\top E[(X-\mu)(X-\mu)^\top]u = E[u^\top (X-\mu)(X-\mu)^\top u] \\ &= E[((X-\mu)^\top u)^\top (X-\mu)^\top u] = E[((X-\mu)^\top u)^2] \geq 0 \end{aligned}$$

so that the covariance matrix is positive semidefinite.

5. $X \in R^n$ and $Y \in R^m$ are independent random vectors. Their expectations and covariance matrices are $E[X] = 0$, $Cov[X] = I$, $E[Y] = \mu$ and $Cov[Y] = \sigma I$, where I is the identity matrix of the appropriate size and σ is scalar. What is the expectation and covariance matrix of the random vector $Z = AX + Y$, where $A \in R^{m \times n}$?

Solution.

The expectation of Z can be obtained from the definition by applying the linearity of expectation,

$$E[Z] = E[AX + Y] = AE[X] + E[Y] = 0 + \mu = \mu.$$

The covariance of Z is $Cov[Z] = E[ZZ^\top] - E[Z]E[Z]^\top = E[ZZ^\top] - \mu\mu^\top$. Substituting the definition of Z, we get the expression below.

$$\begin{aligned} E[ZZ^\top] &= E[(AX+Y)(AX+Y)^\top] = \\ &= E[AXX^\top A^\top + YX^\top A^\top + AXY^\top + YY^\top] = \\ &= AE[XX^\top]A^\top + E[YX^\top]A^\top + AE[XY^\top] + E[YY^\top]. \end{aligned}$$

Here we can substitute $E[XX^\top] = I$ and $E[YY^\top] = \sigma I + \mu\mu^\top$. Because X and Y are independent, $E[XY^\top] = E[X]E[Y]^\top = 0$, similarly $E[YX^\top] = 0$. We get $E[ZZ^\top] = AA^\top + \sigma I + \mu\mu^\top$, therefore $Cov[Z] = AA^\top + \sigma I$.