Theodor Stoican

Data Engineering and Analytics

ID: ge93pas

Matriculation number: 03725732

1.

The basic idea for this exercise is that having *A* of size *m x n* and *C* of size *m x p* in $A \cdot B = C$, then B must of size *n x p*.

Hence we play with the expression in the function according to this idea, by treating each term separately. We know that the final result must be *1x1*. Hence each of the following terms must yield a 1x1 partial result so that we may able to add/subtract them afterwards.

- $x^T Ay$ - ((1 x m) x (m x n) x (n x 1)) - we deduced the dimensions of A so that they could match the other 2. We proceed similarly with the other ones.
- Bx - ((1 x m) x (m x 1))
- $y^T CzD$ - ((1 x n) x (n x p) x (p x q) x (q x 1))
- $y^T E^T y$ - ((1 x n) x (n x n) x (n x 1))
- F - (1 x 1)

2.

Let's expand f(x) to see the pattern in the addition:

$$f(x) = x_1x_1M_{11} + x_1x_2M_{12} + \ldots\ldots + x_2x_1M_{21} + x_2x_2M_{22} + \ldots$$

The intuition is that the last 2 factors in each term resembles the matrix multiplication between x^T and M^T (x_1, x_2, etc. multiplied with M_{11}, M_{12}, etc.). Hence we can safely say that we need this operation: $x^T M^T$. The only question remaining is how do we get the first *x's which appear at the front of each term.*

We notice that $x^T M^T$ has the following size: ((1 x n) x (n x n)) = (1 x n). However, our function f(x) must return a 1 x 1. Intuitively we can think of a vector that transforms what we have so far into a (1 x 1) element and which also adds one 'x' to each term. If we do the math, we'll see that this vector is **x** and that it must be multiplied to the right of the current result in order to yield a 1x1 element after multiplication. Hence:

$$f(x) = x^T M^T x$$

3.

a) $det(A) \neq 0$

b) (Theorem) $det(A) = \prod_i \lambda_i$

Hence we compute the determinant using the eigenvalues:

$det(A) = -5 \cdot 0 \cdot 1 \cdot 1 \cdot 3 = 0$, hence the system does not have a unique solution for any b.

4.

$BA = AB = I$ implies that A is invertible. This, in turn, implies that $det(A) \neq 0$. At the same time we know that $det(A) = \prod_i \lambda_i$. Therefore no eigenvalue can be 0.

5.

"PSD => no negative eigenvalues"

Let $Ax = \lambda x$, according to the definition of eigenvalues.

However, we know that $x^T A x \geq 0$ from the hypothesis. By replacing Ax, we have $x^T \lambda x \geq 0$, which can be rewritten as $\lambda x^T x \geq 0$. We know however that $x^T x \geq 0 \forall x$. Therefore λ must be greater or equal to 0.

"no negative eigenvalues => PSD"

Let $Ax = \lambda x$ according to the definition of eigenvalues. We know that $\lambda \geq 0$ from the hypothesis. By multiplying with x^T to the left, we get:

$x^T A x = x^T \lambda x = \lambda x^T X$. Since $\lambda \geq 0$ and $x^T x \geq 0 \forall x$, then $x^T A x \geq 0$.

Therefore, by proving both implications, PSD <=> no negative eigenvalues.

6.

$$B = A^T A - PSD \, \forall A \text{ iff}$$

$$x^T B x \geq 0 \forall x \text{ iff}$$

$$x^T A^T A x \geq 0 \forall x \text{ iff}$$

$$(Ax)^T (Ax) \geq 0 \forall x \text{ - true,}$$

because the inner product of 2 such vectors (containing real numbers, etc.) is always greater or equal to 0.

7.

a)

i) the shape of the function is a convex parabola (second derivative is positive <=> a > 0)

ii) a = 0, b = 0, meaning that the function is a horizontal line, f(x) = c, and each point is a minimum.

iii) the shape of the function is a concave parabola (second derivative is negative <=> a < 0) or a = 0 (hence, a straight line)

b) The minimum can be found by solving the following equation: $f'(x) = 0$.

$$ax + b = 0$$

$$x = \frac{-b}{a}$$ - closed form expression

8.

a) $\nabla g(x) = \frac{1}{2} \cdot 2 \cdot Ax + b^T$ => $\nabla^2 g(x) = A$. The optimisation problem has unique solution iff A - positive semi-definite (theorem).

b) Intuitively if the function is PSD, then its graph will be convex and will hence have a minimum. If it contains a negative eigenvalue however, then it contains a saddle point too.

c) $\nabla g(x) = 0$ => $Ax + b^T = 0$ => $x = -A^{-1}b^T$

9.

This essentially states that conditional independence implies independence. I will offer a counterexample to disprove the statement.

Let us consider a dice with the following 6 outcomes: {1,2,3,4,5,6}. Now assume that we have 3 events:

- A = {1,2,3} - the event in which one of 1,2, or 3 is on the top face of the dice after rolling it
- B={4,5,6} - the event in which one of 4,5, or 6 is on the top face of the dice after rolling it
- C={4,5} - the event in which one of 4 or 5 is on the top face of the dice after rolling it

Next, let's compute the probabilities stated in the problem statement.

$$P(A|B,C) = \frac{P(A,B|C)}{P(B|C)} = \frac{0}{1} = 0$$ and

$$P(A|C) = 0$$

We notice that in the presence of C, A and B are independent by examining the values (the intuition is that if we already know that the top face of the dice does not contain any number from A by observing C, we don't need any information from B anymore).

At the same time:

$P(A|B) = 0$ and

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

Hence, A and B are not independent(the intuition is that B actually tells us some new information this time, unlike in the previous case).

10.

This essentially states that independence implies conditional independence. I will offer a counterexample to disprove the statement.

Let us consider a dice with the following 6 outcomes: {1,2,3,4,5,6}. Now assume that we have 3 events:

- A = {1,2,3} - the event in which one of 1,2, or 3 is on the top face of the dice after rolling it
- B={1,5} - the event in which one of 1 or 5 is on the top face of the dice after rolling it
- C={1,2,6} - the event in which one of 1, 2, or 6 is on the top face of the dice after rolling it

Next, let's compute the probabilities stated in the problem statement.

$$P(A\,|\,B,C) = \frac{P(A,B\,|\,C)}{P(B\,|\,C)} = \frac{\frac{1}{3}}{\frac{1}{3}} = 0 \text{ and}$$

$$P(A\,|\,C) = \frac{2}{6} = \frac{1}{3}$$

We notice that in the presence of C, A and B are not independent by examining the values (the intuition is that if we know that the top face of the dice contains numbers from C, then by observing B as well, it means that on the top face there is a number from the intersection of B and C - hence a number from A).

At the same time:

$$P(A\,|\,B) = \frac{P(A,B)}{P(B)} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2} \text{ and}$$

$$P(A) = \frac{1}{2}$$

Hence, A and B are independent(the intuition is that observing B whose correspondent set contains one number from A and one outside of A, we don't gain any new insight).

11.

1. $P(a) = \int_b \int_c P(a,b,c)\,dc\,db$

2. $P(c\,|\,a,b) = \dfrac{P(c,a\,|\,b)}{P(a\,|\,b)} = \dfrac{\frac{P(a,b,c)}{P(b)}}{P(a\,|\,b)} = \dfrac{P(a,b,c)}{P(a\,|\,b)\cdot P(b)} = \dfrac{P(a,b,c)}{P(a,b)} = \dfrac{P(a,b,c)}{\int_c P(a,b,c)\,dc}$

3. $P(b\,|\,c) = \dfrac{P(b,c)}{p(c)} = \dfrac{\int_a P(a,b,c)\,da}{\int_a \int_b p(a,b,c)\,db\,da}$

12.

We can translate the problem into probabilities:

$$P(pos\,|\,sick) = \frac{95}{100}$$

$$P(\neg pos\,|\,\neg sick) = \frac{95}{100}$$

$$P(sick) = \frac{1}{1000}$$

$$P(sick\,|\,pos) = \,?$$

By applying Bayes' rule:

$$P(sick\,|\,pos) = \frac{P(pos\,|\,sick)\cdot P(sick)}{P(pos)} = \frac{P(pos\,|\,sick)\cdot P(sick)}{P(pos\,|\,sick)\cdot P(sick) + P(pos\,|\,\neg sick)\cdot P(\neg sick)} = \frac{\frac{95\cdot 1}{100\cdot 1000}}{\frac{95\cdot 1}{100\cdot 1000} + \frac{5\cdot 999}{100\cdot 1000}}$$

$$= 0.01866404715$$

13.

13.

$$E[f(x)] = E[ax + bx^2 + c] = E[ax] + E[bx^2] + c =$$

$$= aE[x] + bE[x^2] + c$$

$$E[x] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \quad \Bigg| \quad t = \frac{x-\mu}{\sqrt{2}\,\sigma}, \quad dt = \frac{1}{\sigma\sqrt{2}} \, dx$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\,\sigma t + \mu) e^{-t^2} \, dt$$

$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\,\sigma \int_{-\infty}^{\infty} t e^{-t^2} \, dt + \mu \int_{-\infty}^{\infty} e^{-t^2} \, dt \right)$$

$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\,\sigma \left[-\frac{1}{2} e^{-t^2} \right]_{-\infty}^{\infty} + \mu\sqrt{\pi} \right)$$

$$= -\frac{\sqrt{2}\,\sigma}{2\sqrt{\pi}} \underbrace{\left(e^{-\infty} - e^{-\infty} \right)}_{0} + \mu = \mu$$

For computing $E[x^2]$, we can either use the same technique or make use of the Gaussian variance.

$$Var(x) = E[x^2] - E[x]^2$$

$$\Rightarrow E[x^2] = Var(x) + E[x]^2$$

$$= \sigma^2 + \mu^2$$

$$\Rightarrow E[f(x)] = a\mu + b(\sigma^2 + \mu^2) + c$$

14.

- $E[g(x)] = E[Ax] = A\,E[x] = A\mu$

- $E[g(x)\,g(x)^T] = E[A x x^T A^T] = A\,E[x x^T]\,A^T$

$E[xx^T] = E[(x-\mu)(x-\mu)^T - \mu\mu^T + x\mu^T + \mu x^T] =$

$= \Sigma - \mu\mu^T + E[x]\mu^T + \mu E[x^T] =$

$= \Sigma \cancel{-\mu\mu^T} \cancel{+\mu\mu^T} + \mu\mu^T$

$= \Sigma + \mu\mu^T$

- $E[g(x)^T g(x)] = E[x^T \underbrace{A^T A}_{B} x]$ B (notation)

$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$ (since this is a 1x1 matrix)

$\Rightarrow E[x^T A x] = E\left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j\right]$

$= a_{ij}\, E[x_i x_j]$ (Einstein summat

$= a_{ij}(\sigma_{ij} + \mu_i \mu_j)$

$= a_{ij}\sigma_{ij} + a_{ij}\mu_i\mu_j$

Σ is symmetric

$= a_{ij}\sigma_{j,i} + \mu^T A \mu$ → similar to the beginning

$= \sum_{i=1}^{n} [A\Sigma]_{i,i} + \mu^T A \mu = tr(A\Sigma) + \mu^T A \mu$

$\Rightarrow$ Our value can be obtained by substituting A with A^TA.

$$\rightarrow E\left[g(x)^T g(x)\right] = tr\left(A^T A \Sigma\right) + \mu^T A^T A \mu$$

- $Cov\left[g(x)\right] = Cov\left[Ax\right] = E\left[Ax \cdot x^T A^T\right] - E\left[Ax\right] \cdot E\left[Ax\right]^T =$

$= \Sigma + \mu\mu^T - A\mu\mu^T A^T$ (computed during the previous points)