CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

Quiz (Closed Book)

Roll Number: _____

Name: _____

For all questions, if you feel that some information is missing, make justifiable assumptions, state them clearly, and answer the question.

Relevant Formulae

• Poisson: $P(k|\lambda) := \lambda^k \exp(-\lambda)/(k!)$

• Exponential: $P(x;\lambda) = \lambda \exp(-\lambda x); \forall x > 0$

• Gamma:

$$P(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}$$

• Gamma function: $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx$ for real-valued z.
When z is integer valued, then $\Gamma(z) = (z-1)!$, where ! denotes factorial.
For all z, $\Gamma(z+1) = z\Gamma(z)$.

• Univariate Gaussian:

$$P(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\frac{(x-\mu)^2}{\sigma^2}\right)$$

• Product of two univariate Gaussians: $G(z;\mu_1,\sigma_1^2)G(z;\mu_2,\sigma_2^2) \propto G(z;\mu_3,\sigma_3^2)$
where

$$\mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}; \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

• Multivariate Gaussian:

$$G(x;\mu,C) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(x-\mu)^\top C^{-1}(x-\mu))$$

• $d(Ax) = Adx$
• $d(x^\top Ax) = x^\top(A+A^\top)dx$

---

1. (15 points)

   Consider a dataset $\{x_n \in \mathbb{R}^D\}_{n=1}^N$ where $D > 3$ and $\sum_{n=1}^N x_n = \mathbf{0}$.

   • [2 points] Propose a quantitative measure of the quality of fit of any given 1D line (passing through the origin) in $\mathbb{R}^D$ to the dataset.

   • [5 points] Derive and propose an implementable algorithm for finding the representation of the best-fitting line described in the previous sub-question.

   • [3 points] Propose a quantitative measure of the quality of fit of any given 2D plane (passing through the origin) in $\mathbb{R}^D$ to the dataset.

   • [5 points] Derive and propose an implementable algorithm for finding the representation of the best-fitting 2D plane described in the previous sub-question.

   Measure of fit can be the sum of squared distances between datum and its projection on the line (or plane), or the "reconstruction error"

   Minimizing the reconstruction error is equivalent to maximizing the variance of the data on the projected line/plane (as we've discussed)

   After this point, it just boils down to PCA analysis that we've covered in the slides.

   Please check the lecture slides.

   ---

2. (10 points)

   Consider a classification application, involving 2 classes, where the distribution of each class is modeled by a multivariate Gaussian.

   • [3 points] State the most general condition when the decision boundary underlying the classification problem is linear.

   • [7 points] Derive the equation of this decision boundary.

   Please check the lecture slides.

   ---

3. (15 points)

   Consider a matrix A of size $M \times N$ of rank M, where $2 < M < N$. Consider matrices $B := A^\top A$ and $C := AA^\top$.

   • [1.5 points: 0.5 + 0.5 + 0.5] Which of A, B, and C have a singular value decomposition ?

   All of them. Please check the lecture slides.

   • [1.5 points: 0.5 + 0.5 + 0.5] Which of A, B, and C have an eigen decomposition ?

   Only the symmetric square matrices B and C. Please check the lecture slides.

   • [6 points: 2 + 2 + 2] Derive relationships between the eigenvalues and/or singular values of the matrices A, B, and C.

   If SVD of $A = USV^\top$

(where U is orthogonal of size $M \times M$, S is rectangular diagonal of size $M \times N$ with M non-zero entries on the diagonal, and V is orthogonal of size $N \times N$),

then $B = VS^\top SV^\top$ and $C = USS^\top U^\top$.

For B, the matrix $S^\top S$ is of size $N \times N$ with only $M < N$ non-zero entries on the diagonal.

For C, the matrix $SS^\top$ is of size $M \times M$ with full rank.

Eigenvectors of B are columns of V. Eigenvectors of C are columns of U.

Eigenvalues of B and C are the diagonal entries in $S^\top S$ or $SS^\top$, respectively, which are either the square of the underlying singular values of A or zero.

• [6 points] Prove that any real symmetric matrix has real eigenvalues.

Please check the lecture slides.

---

4. (15 points)

Let continuous random variables X and Y be independent and have Gaussian distributions, respectively, $G(0, 1)$ and $G(2, 3)$.

For all questions below, give a simple analytical expression, e.g., without any integrals or derivatives.

• [3 points] Derive the distribution of the random variable $A := X^2$.

Please check the lecture slides.

• [3 points] Derive the distribution of the bivariate random variable $B := (X, X)$.

When $Z_1 = Z_2 = X$, then $P(Z_1, Z_2) = G(X; 0, 1)$

When $Z_1 \neq Z_2$, then $P(Z_1, Z_2) = 0$

• [3 points] Derive the distribution of the bivariate random variable $C := (X, Y)$.

Because of independence, $P(X, Y) = G(X; 0, 1)G(Y; 2, 3)$

• [3 points] Derive the distributions of the random variable $D := Y + Y$.

$D = 2Y$

Using transformation of random variables, D is a Gaussian with mean 4 and variance 12.

• [3 points: 1.5 + 1.5] Give expressions for the distributions of the random variables $E := Y - X$ and $F := Y + X$.

X and $-X$ have the same PDF (standard normal).

So, E and F will have the same PDF.

`https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables#Independent_random_variables`

Because X and Y are independent, the PDFs of E and F will be Gaussian with mean 2 and variance 4.
