# CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

## End-Semester Examination (Maximum Points 85; Closed Book)

Date: 14 Nov 2018. Time: 2 pm - 5 pm

**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**

• Univariate Gaussian: $G(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

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

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

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

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

• 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.

• $KL(P\|Q) = \int_x P(x)\log(P(x)/Q(x))dx$

---

1. [20 points]

   (a) (2 points) For what kind of estimators, biased or unbiased, does the Cramer-Rao lower bound theorem apply ?

   ---

   Unbiased estimators

   ---

   (b) (2 points) For what kind of estimators, biased or unbiased, does the Bayesian Cramer-Rao lower bound theorm apply ?

   ---

   Both

   ---

   (c) (4 points) Is the maximum-likelihood estimator of the univariate Gaussian mean (when variance is known) an efficient estimator ? Prove or disprove.

When $X_1, X_2, \cdots, X_N$ are independent, variance of sample mean is sum of variances of $X_n/N$ $= \sum_{n=1}^{N} \sigma^2/N^2 = \sigma^2/N$

CRLB gives the minimum variance as the inverse of Fisher information, which also evaluates to $\sigma^2/N$

(d) (8 points: 2 + 2 + 2 + 2) Suppose you are given two $N$-sized data samples $\{x_n\}_{n=1}^{N}$ and $\{y_n\}_{n=1}^{N}$. It is known that each sample point $x_n$ is drawn independently from a Gaussian with mean $\mu_X$ and variance $\sigma^2$. It is known that each sample point $y_n$ is drawn independently from a Gaussian with mean $\mu_Y$ and variance $\sigma^2$. For a fixed finite sample size $N$, suppose you define a statistic that is the distance between the sample-mean estimates, i.e., $|\widehat{\mu}_X - \widehat{\mu}_Y|$.

• Give a clear and precise way to estimate the distribution / histogram of $|\widehat{\mu}_X - \widehat{\mu}_Y|$, over different samples, for the fixed finite sample size $N$, *when* $\mu_X = \mu_Y$.

We can estimate this distribution empirically

We know $N$, $\mu_X$, $\mu_Y = \mu_X$, $\sigma^2$

Draw $N$-sized samples $\{x_n\}_{n=1}^{N}$ and $\{y_n\}_{n=1}^{N}$. Compute the test statistic. Repeat this experiment, say, thousands of times (more the better). Compute the histogram of the test statistics, obtained over all experiments.

• How can you use the aforementioned histogram to test if two $N$-sized samples have been drawn from Gaussians with the same mean or different means ? Note: you are given that the variances of the two Gaussian are equal to $\sigma^2$.

If $\mu_X$ and $\mu_Y$ are far apart, then the test statistic will take a value that will be atypical under the learned distribution (in the previous question). Thus, we can use the previous distribution to look up the value of the CDF at the observed test statistic: if the CDF value is very small (say, less than a threshold of 5%) or very large (say, more than a threshold 95%), then we tend to fail to accept that the means are the same. Closer the CDF value to 0 or 100, stronger is our belief in the means being different.

• Can you use the aforementioned histogram to test if two $M$-sized, where $M \neq N$, samples have been drawn from Gaussians with the same mean or different means ? Note: you are given that the variances of the two Gaussians are equal to $\sigma^2$. Argue why or why not.

No, because the histogram of the test statistic depends heavily on the sample sizes.

• Can you use the aforementioned histogram to test if two $N$-sized, samples have been drawn from Gaussians with the same mean or different means, when you are told that variances of the two Gaussians are equal to $\alpha^2$, where $\alpha \neq \sigma$ ? Argue why or why not.

No, because the histogram of the test statistic depends heavily on the sample sizes.

(e) (4 points: 2 + 2) Consider a Gaussian *mixture* probability density function (PDF) $P(X) := \sum_{n=1}^{N} w_n G(X; \mu_n, \sigma_n^2)$.

• Derive an expression for the mean of $P(X)$.

• Derive an expression for the variance of $P(X)$.

```
https://en.wikipedia.org/wiki/Mixture_distribution#Moments
```

---

2. [5 points] Consider Bayesian inference using the posterior probability density function $P(\theta)$ on $\theta \in \mathbb{R}$, using which you want to infer a value $\widehat{\theta} \in \mathbb{R}$ using a specific loss function deined as follows:

   $L(\widehat{\theta}|\theta) := w(\theta - \widehat{\theta})$, when $\theta \geq \widehat{\theta}$

   $L(\widehat{\theta}|\theta) := (1 - w)(\widehat{\theta} - \theta)$, when $\widehat{\theta} > \theta$

   Here, $w$ is a fixed specified real-valued scalar and $w \in [0, 1]$.

   For this loss function,

   - Define the risk function.

   ---

   Please see class notes.

   ---

   - Derive the value $\widehat{\theta}$ that minimizes the risk function ? Specify $\widehat{\theta}$ in the simplest possible terms, without including any expectations.

   ---

   Follow the derivation of the case when $w = 0.5$, when the $\widehat{\theta}$ is the median.

   When $w \neq 0.5$, the estimate $\widehat{\theta}$ is the $(100w)$-th percentile.

   ---

3. [10 points] Suppose you want to build a classifier to seperate (i) pictures of human faces (say, passport-size photographs) from (ii) pictures (same size images as that of human faces) of faces of other living beings in the entire universe. The class of other living beings includes animals and birds on the earth and living beings (unseen) in other parts of the universe. You know that the distribution of human-face images is multivariate Gaussian. You cannot assume anything about faces that you haven't seen, i.e., have *no* data. Describe clearly and precisely an implementable algorithm for:

   - (5 points) Training / learning a probabilistic model that will act as the classifier.

   ---

   We can only rely on the data observed for the human-face class. The non-human-face class PDF cannot be modeled, because we haven't seen any aliens and we don't have all data / information about the other class.

   Hence, we learn a PDF for the human-face class, by fitting a multivariate Gaussian. State this procedure.

   We will use this PDF to evaluate the probability density of a new datum to being a human face. A low probability density indicates a non-human face.

   ---

   - (5 points) Applying the learned model to classify an image as human or non-human.

   ---

   See the above answer. State the formula for evaluating the PDF.

   ---

4. [15 points] Suppose you want to classify a set $\mathcal{S}$ of high-quality images of handwritten digits in some script (0 to 9). Each image has $N$ pixels and, thus, the space of representation of these images is $\mathbb{R}^N$. You are told that all images for any particular digit lie in some *hyperplanar subspace* $\mathbb{R}^D$ of $\mathbb{R}^N$, where $D \ll N$. Different digits correspond to different *non-intersecting* subspaces of possibly *different dimensions* $D$. You aren't told anything about the distribution of the images of the digit within its subspace and you are forbidden to model the within-subspace distribution.

You are given a sufficiently large *training set* $\mathcal{T}$ of high-quality images of digits, with labels indicating the digit corresponding to the image. The set $\mathcal{S}$ isn't associated with such labels.

You must use all the aforementioned information (without any additional assumptions on the distributions of images) to solve the following questions.

- (8 points) Design an algorithm to classify each image in the set $\mathcal{S}$ as even or odd.

---

`https://en.wikipedia.org/wiki/Hyperplane`

Perform PCA no each digit class in the training set. Select the subspace / hyperplane, passing through the mean, corresponding to eigenvectors for all non-zero (positive) eigenvalues (each subspace has its own coordinate system, containing an origin.) You'll have one such subspace for each digit.

For each image in the test set, find the distance between the image and the subspace corresponding to each digit. For one such digit, the distance will be zero. Classify the image into that digit's class.

---

- (7 points: 5 + 2) Now, assume that the images in another set $\mathcal{U}$ also need to be classified, but this image data is *corrupted with measurement errors* such that an independent random standard-normal perturbation gets incorporated into the measurement at each pixel. Design two PDFs $P(\text{image}|\text{even})$ and $P(\text{image}|\text{odd})$ that indicate probability densities of any image coming from the even set and the odd set, respectively. Use the PDFs designed in the previous part of this question to assign a probability to an image in $\mathcal{U}$ of being even (or odd).

---

For each digit class, design a PDF where the probability density for an image is a (univariate) Gaussian based on (i) the (scalar) distance of the image from the hyperplane and (ii) the variance as the noise variance.

Then, $P(\text{image}|\text{even}) = \sum_{i=0,2,4,6,8} P(\text{image}|\text{digit} = i)$

$P(\text{image}|\text{odd}) = \sum_{i=1,3,5,7,9} P(\text{image}|\text{digit} = i)$

To classify, $P(\text{odd}|\text{image}) = P(\text{image}|\text{odd})/(P(\text{image}|\text{odd}) + P(\text{image}|\text{even}))$

---

5. [15 points] Consider a univariate real-valued random variable $X$ with an associated probability density function (PDF) $P(X)$ having mean $\mu$, median $m$, and standard deviation $\sigma$. In each of the following questions, you will be given two expressions. For each pair of expressions, prove or disprove (i) if the expressions are equal, (ii) if the first expression takes values always less than (or $\leq$) the other, (iii) the first expression takes values greater than (or $\geq$) the other, or (iv) there isn't any such relationship between the expressions. You can use theorems derived in class, but state all theorems and arguments in the proof clearly.

- (4 points) Expressions $|E[X - m]|$ and $E[|X - m|]$, where $|\cdot|$ is the absolute-value function.

---

Jensen's inequality: $f(E[Y]) \leq E[f(Y)]$ for convex $f$. Absolute-value function is convex.

---

- (4 points) Expressions $E[|X - \mu|]$ and $\sqrt{E[(X - \mu)^2]}$.

---

Jensen's inequality: $f(E[Y]) \geq E[f(Y)]$ for concave $f$. Square-root function is concave on the positive-real line.

---

- (4 points) Expressions $E[|X - m|]$ and $E[|X - \mu|]$.

---

We proved in class that the median $m$ minimizes $E[|X-a|]$, over all possible $a$. So, any other $a \neq m$ can only be equal or larger than $E[|X-m|]$.

---

- (3 points) Expressions $|\mu - m|$ and $\sigma$.

---

Because of all the previous results, $|\mu - m| \leq \sigma$.

---

6. [10 points] Suppose you are given a data sample $\{x_n\}_{n=1}^{N}$, where each $x_n$ is drawn from a probability density function (PDF) that is multivariate Gaussian. You want to build a sampler for generating more observations $x$ from that same PDF. Describe, and clearly justify, an algorithm for the sampler.

---

Let $X := \mu + AW$, as per the definition of the multivariate Gaussian (see class notes)

Perform PCA on the data. Get estimates of $\mu$ and $C$

Perform eigendecomposition on $C = V \Lambda V^\top$ to get $A = V\sqrt{\Lambda}$

Then, generate new sample points $x = \mu + Aw$, where $w$ is drawn from i.i.d. standard-normal PDFs (see class notes)

---

7. [10 points] Consider two random variables $U_1$ and $U_2$ that are independent, with each having a uniform distribution over $(0,1)$. Consider the two transformed random variables $Z_1 := \sqrt{-2\log U_1}\cos(2\pi U_2)$ and $Z_2 := \sqrt{-2\log U_1}\sin(2\pi U_2)$. Using the theory of transformation of random variables, (i) derive the joint probability density function (PDF) for the bivariate random variable $Z := (Z_1, Z_2)$ and (ii) derive the marginal PDFs $P(Z_1)$ and $P(Z_2)$.

---

$U_1 = \exp(-0.5(Z_1^2 + Z_2^2))$

$U_2 = (0.5/\pi)\arctan(Z_2/Z_1)$

Consider the tranformation $(Z_1, Z_2) = f(U_1, U_2)$

Then, $Q(Z_1, Z_2) = P_{U_1,U_2}(f^{-1}(Z_1, Z_2))|\det(J)|$, where $J$ is the Jacobian matrix associated with $f^{-1}(\cdot,\cdot)$.

We know that $P_{U_1,U_2} = 1$ and

that $|\det(J)| = (0.5/\pi)\exp(-0.5(Z_1^2 + Z_2^2)) = \sqrt{0.5/\pi}\exp(-0.5Z_1^2)\sqrt{0.5/\pi}\exp(-0.5Z_2^2)$

Thus, $Z_1$ and $Z_2$ are two independent standard-normal random variables

This transformation is known as the Box-Muller transform.

---