CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

End-Semester Examination (Maximum Points 105)

Date: 16 Nov 2017. Time: 5:30 pm - 8:30 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: $P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

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

• 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) (4 points) Maximum-likelihood estimators are always unbiased ? Prove or disprove.

   The ML estimator for the variance of a univariate Gaussian is biased. This is easy to show.

   (b) (3 points) Suppose you are performing principal component analysis (PCA) on data from several individuals, where the data from individual i involves D observations $x_i \in \mathbb{R}^D$, e.g., the individual's height, weight, blood cell counts, etc.

   (i) (1 point) Can changing the units of one of the measurements, say, height from meters to centimeters (i.e., the height changing from 1 to 100) modify the first principal component, if the other measurements' units remained unchanged ?

Yes. For instance, if the original eigenvalues were identical and eigenvectors non-unique, then scaling up one dimension will make the principal eigenvector correspond to that direction.

(ii) (2 points) What is a (standard) technique to make units of different measurements commensurate in PCA ? Clearly state the algorithm and justify it briefly.

Standardize each variable, i.e., subtract mean and scale by the standard deviation.

(c) (4 points) Suppose you are performing principal component analysis (PCA) on data $\{x_i\}_{i=1}^{I}$ whose mean is non-zero. Suppose you compute the covariance without subtracting the mean from (i.e., centering) the data.

(i) (2 points) Will this affect the computed principal components /directions ? If so, explain clearly or illustrate clearly with a picture. If not, prove it.

Yes. If mean was far from origin, then the principal eigenvector now points (roughly) towards the mean.

(ii) (2 points) Will this affect the computed variances along the principal components ? If so, explain clearly or illustrate with a picture. If not, prove it.

Yes. If mean was far from origin, first eigenvalue will be large.

(d) (4 points) Suppose you are performing principal component analysis (PCA) on data from several individuals, where the data from individual i involves D observations $x_i \in \mathbb{R}^D$.

(i) (2 points) Suppose the number of individuals is $N < D$. In this case, what can you say about the eigenvalues of the covariance matrix ?

The last $D - N - 1$ eigenvalues will be exactly zero. e.g., for 2 points in a 3D space, the last 2 eigenvalues will be zero.

(ii) (2 points) Suppose, during PCA, you find that all eigenvalues of the covariance matrix are identical and non-zero ? What does this tell us about the data in the context of its (i) scatter plot, (ii) modes of variation, and (iii) relationships between the variables?

Then, $C = Q\Lambda Q^\top = Q(\lambda I)Q^\top = \lambda Q I Q^\top = \lambda I$

(i) Isotropic.

(ii) Non-unique.

(iii) Uncorrelated (not necessarily inpendenent), in general.

Independent, if we assume the joint / mulivariate PDF $P(X_1, X_2, \cdots)$ was normal, which can then be factored as $P(X_1)P(X_2)\cdots$.

(e) (5 points) Suppose you want to evaluate the Kullback-Leibler divergence between distributions $P(X)$ and $Q(X)$ in a real-world application, where X takes values in a high-dimensional space. Assume $P(X)$ and $Q(X)$ have support over the entire domain. However, (1) the underlying integral is analytically intractable and (2) numerical approximation of the integral via a Riemann sum (i.e., finite sum approximation to the area under the "curve") is also computationally infeasible because of the large dimensionality involved.

(i) (3 points) In such a case, describe a computationally-efficient statistical method to approximate the value of the integral. The method must allow estimation of the integral value up to any arbitrary level of accuracy. Make reasonable assumptions on the real-world distributions.

$KL(P\|Q) = \int_x P(x)\log(P(x)/Q(x))dx = E_{P(X)}[\log(P(X)/Q(X))]$
$\approx \sum_{x_i \sim P(X)} \log(P(x_i)/Q(x_i))$

(ii) (2 points) Describe under what circumstances will the statistical method be significantly more efficient over a Riemann-sum approximation.

When the distribution has mass that is almost-all concentrated to a subspace (e.g., 1D subspace in a 100 dimensional space), because Riemann-sum approximation in a 100-D subspace will be extremely costly.

2. (10 points) Consider a bivariate Gaussian random variable $X := [X_1, X_2]^\top$ with mean $\mu := [\mu_1, \mu_2]^\top$ and covariance matrix C with the element in row i and column j represented as C_{ij}.

(i) (4 points) What are the marginal probability density functions of univariate random variables X_1 and X_2 ? Derive the expressions solely in terms of the components of μ and C.

See next question; general multivariate case.

(ii) (6 points) What is the conditional probability density function $P(X_1|X_2 = a)$? Derive the expression solely in terms of the components of μ and C.

`en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case`

Proof:

Consider the elements of the covariance matrix as $C_{11} = \sigma^2$, $C_{22} = \sigma^2$, $C_{12} = C_{21} = \rho\sigma_1\sigma_2$

Use the definition of conditional probability and use completion of squares.

`en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions`

3. (15 points) Consider a multivariate Gaussian in D dimensions modeled by random vector $X := [X_1, X_2, \cdots, X_D]^\top$ with mean μ and covariance matrix $C := AA^\top$.

   (i) (5 points) Derive the expression of the marginal probability density function of the set of $K < D$ random variables $Y := [X_1, X_2, \cdots, X_K]^\top$, in terms of the components of mean μ and covariance C.

   `en.wikipedia.org/wiki/Multivariate_normal_distribution#Marginal_distributions`

   Proof:

   We know that $X = AW + \mu$

   Consider i-th row of A as $a_i^\top$.

   Consider a projection matrix B such that $Y = BAW + B\mu$.

   $B\mu$ extracts the top K elements of μ. This the new mean.

   BA extracts the top K rows of A.

   Thus, the new covariance of Y is $BA(BA)^\top$ that is the inner products of rows $a_i^\top$, which is the top left of C.

   (ii) (5 points) What is the mean and the covariance matrix of random vector Y ?

   See the solution above.

   (iii) (5 points) What is the conditional probability density function $P(X_1|X_2 = a)$ when covariance matrix C is diagonal ? Derive the expression in terms of the components of μ and C.

   The bivariate PDF $P(X_1, X_2)$ is a bivariate Gaussian with mean and covariance involving appropriate subsets of numbers from μ and C.

   The bivariate PDF gets factored when C is diagonal. We know that the variables are uncorrelated, i.e., $\rho = 0$.

   `en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions`

4. (10 points) Suppose we model a n-sized data sample as $\mathbf{X_n} := \{X_1, X_2, \cdots, X_n\}$ where the i-th observation X_i is an n-vector $X_i := [X_{i1}, \cdots, X_{in}]^\top$. We know that X_{ij} and X_{ik} are independent (for any j, k) and all $X_{i\bullet}$ have the normal probability density function $G(\mu_i, \sigma^2)$. We want to estimate parameters $\{\mu_i\}_{i=1}^n$ and σ^2 given data $\{x_1, \cdots, x_n\}$. This is a model where the number of parameters (i.e., $n + 1$) increase with sample size (i.e., n). Such situations indeed arise in real-world applications and have been well studied in the literature.

   (i) (5 points) Derive the maximum-likelihood estimators for the parameters.

   (ii) (5 points) As the sample size n tends to ∞, do the ML estimates converge to the true values ?

   The ML estimates are:

   $\widehat{\mu}_i = (1/n) \sum_{j=1}^n X_{ij}$

$\widehat{\sigma}^2 = (1/n^2) \sum_{i=1}^n \sum_{j=1}^n (X_{ij} - \widehat{\mu}_i)^2$

Yes, the ML estimates converge to the true values as $n \to \infty$

The mean converges because of the law of large numbers.

The variance also converges because for each i, $(1/n) \sum_j (X_{ij} - \widehat{\mu}_i)^2$ can be rewriten as $\sum_j (1/n) X_{ij}^2 + \sum_j (1/n) \widehat{\mu}_i^2 - \sum_j (1/n) 2 X_{ij} \widehat{\mu}_i$, for each i

For each i, the law of large numbers says that (i) the 1st term converges to the variance + μ_i^2, (ii) 2nd term tends to μ_i^2, (iii) 3rd term tends to $-2\mu_i^2$.

So, the sum of the three terms tends to the variance σ^2.

5. (20 points) Suppose we model a n-sized data sample as $\mathbf{X_n} := \{X_1, X_2, \cdots, X_n\}$ where the i-th observation X_i is a 2-vector $X_i := [X_{i1}, X_{i2}]^\top$. We know that X_{i1} and X_{i2} are independent and both have the normal probability density function $G(\mu_i, \sigma^2)$. We want to estimate parameters $\{\mu_i\}_{i=1}^n$ and σ^2 given data $\{x_1, \cdots, x_n\}$. This is a model where the number of parameters $(n+1)$ increase with sample size (n).

   (i) (5 points) Derive the maximum-likelihood estimators for the parameters.

   (ii) (5 points) As the sample size n tends to ∞, do the ML estimates converge to the true values ?

$\widehat{\mu}_i = (1/2) \sum_{j=1}^2 X_{ij}$

Expected value of $\widehat{\mu}_i$, as $n \to \infty$, is $\widehat{\mu}_i$ itself that isn't a function of n.

—

$\widehat{\sigma}^2 = (0.5/n) \sum_{i=1}^n \sum_{j=1}^2 (X_{ij} - \widehat{\mu}_i)^2$

which is equivalent to $\widehat{\sigma}^2 = (0.25/n) \sum_{i=1}^n (X_{i1} - X_{i2})^2$

Now, $X_{i1} - X_{i2}$ is a Gaussian with mean 0 and variance $2\sigma^2$.

Thus, the expectation of $(X_{i1} - X_{i2})^2/n$ is the variance $2\sigma^2$.

Thus, $\widehat{\sigma}^2 \to \sigma^2/2$

—

NO, the ML estimates, for both mean and variance, don't converge to the true values as $n \to \infty$

(iii) (10 points: 8 + 2) Now consider that real problem is to estimate the parameter σ^2 and interpret μ_i as a nuisance variable that models an individual-specific bias in the observations (x_i, y_i) for individual i. You decide to use a Bayesian approach with some prior on the variance $P(\sigma^2)$; don't explicitly assume any particular form for $P(\cdot)$. The Bayesian approach will estimate σ^2 by maximizing the posterior $P(\sigma^2|\mathbf{x_n})$ that is obtained by integrating out the nuisance variables $\{\mu_i\}_{i=1}^n$ from $P(\sigma^2, \{\mu_i\}_{i=1}^n|\mathbf{x_n})$.

- Derive an expression for the Bayes estimate of the variance maximizing posterior $P(\sigma^2|\mathbf{x_n})$.
- As the sample size n tends to ∞, does the Bayes estimate converge to the true σ^2 ?

The proof relies on completion of squares and some algebraic simplification to show that:

$\widehat{\sigma}^2 = (0.5/n) \sum_{i=1}^n (X_{i1} - X_{i2})^2$ that is, desirably, twice as large as the ML estimate.

So, the Bayes estimate, for the variance, converges to the true value as $n \to \infty$

`https://telescoper.files.wordpress.com/2016/11/neyman-scott1.pdf`

http://numerical.recipes/whp/notes/NeymanScott.pdf

6. (10 points)

   (i) (5 points) Find the Kullback-Leibler divergence KL$(G_1 \| G_2)$ between two Gaussian probability density function given by $G_1(x; \mu_1, \sigma_1^2)$ and $G_2(x; \mu_2, \sigma_2^2)$.

   Starting from the definition, the integral evaluates to

   $\log(\sigma_2/\sigma_1) + (\sigma_1^2 + (\mu_1 - \mu_2)^2)/(2\sigma_2^2) - 0.5$

   (ii) (5 points) In Bayesian statistics, for a prior $P(\theta)$ and posterior $Q(\theta)$, the Kullback-Leibler divergence between the prior distribution and the posterior, i.e., KL$(P \| Q)$, is termed as the "surprise".

   • Find an expression for the surprise in the context of Bayesian estimation of a Gaussian's mean parameter μ, when the Gaussian's variance σ^2 is known, the observed sample size is N, the sample mean of the observed data is $\overline{m}$, and the prior on the mean is $G_1(x; \mu_1, \sigma_1^2)$.

   We know what the posterior going to be. See class notes.

   Plug the prior and posterior parameters in the formula above.

7. (10 points) Consider an experiment about tossing a coin N times, with the probability of obtaining a head being μ where $0 \leq \mu \leq 1$. Consider a Gaussian prior on μ with the probability density function $G(\mu; 0.5, \sigma^2)$. Suppose in $N = 3$ flips, we observe $N = 3$ heads. Derive an expression for the maximum-a-posteriori estimate for the parameter μ, as function of σ. Evaluate that expression (approximately; upto second place of decimal) when $\sigma = 0.1$.

   Close to 0.55

   http://www.cs.tut.fi/~hehu/SSP/lecture10.pdf (Pages 22-25)

8. (10 points) Consider the prior density $P(\theta)$ that is known to be conjugate to the likelihood $L(\theta)$. Suppose you need to estimate θ. The prior density $P(\theta)$ is unimodal but the prior information you have indicates a distribution that is multimodal. Design a prior density $Q(\theta)$, using functions $P(\cdot)$, which allows you to model multimodal densities without losing the ease of parameter estimation with conjugate priors. Ensure that $Q(\theta)$ is also a conjugate prior.

   Let $P(\theta)$ be parametrized by parameters $alpha$.

   Define conjugate prior as the so-called "mixture model" as $Q(\theta) := \sum_{i=1}^{K} w_k P(\theta; \alpha_k)$, where $w_k \geq 0$ and $\sum_k w_k = 1$.

   Then, the posterior is a convex weighted combination of the $P(\theta; \alpha_k)L(\theta)$ that we know is also in the same family as $P(\theta)$ (because P is conjugate to L). This mixture-model analytical form for the new posterior is easily obtained thanks to the conjugacy of P and L.

   Thus, the mixture prior Q is also conjugate to the L.

The Bayes-estimate *mean* of the mixture posterior can be easily obtained as a weighted combination of the means of the posterior mixture components.

Of course, the posterior mode and the median aren't as easily computed for the Q prior as compared to the P prior.