Co902 problem sheet, solutions

1. (a) (*Markov inequality*) Let X be a continuous, non-negative random variable (RV) and a a positive constant. Show that:

$$P(X \geq a) \quad \leq \quad \frac{\mathbb{E}[X]}{a}$$

Solution: Let $p(x)$ represent the pdf of RV X. Then:

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty x\, p(x)\, \mathrm{d}x \\ &= \int_0^a x\, p(x)\, \mathrm{d}x + \int_a^\infty x\, p(x)\, \mathrm{d}x \end{aligned}$$

Since $p(x)$ is a pdf, it is everywhere non-negative, so the first term on the RHS must be non-negative. This means:

$$\mathbb{E}[X] \quad \geq \quad \int_a^\infty x\, p(x)\, \mathrm{d}x$$

Since a is the lower bound on the integral above, we can write

$$\int_a^\infty x\, p(x)\, \mathrm{d}x \quad \geq \quad \int_a^\infty a\, p(x)\, \mathrm{d}x$$

which gives

$$\begin{aligned} \mathbb{E}[X] &\geq \int_a^\infty a\, p(x)\, \mathrm{d}x \\ &= a \int_a^\infty p(x)\, \mathrm{d}x \\ &= a\, P(X \geq a) \end{aligned}$$

from which the required result follows.

(b) (*Chebyshev inequality*) Let X be any RV with mean μ_X and whose variance σ_X^2 exists. Show that for any positive constant a:

$$P(|X - \mu_X| \geq a) \quad \leq \quad \frac{\sigma_X^2}{a^2}$$

Solution: First, note that

$$P(|X - \mu_X| \geq a) \quad = \quad P((X - \mu_X)^2 \geq a^2)$$

Here, $(X - \mu_X)^2$ is a non-negative RV. Using the Markov inequality, we get:

$$\begin{aligned} P((X - \mu_X)^2 \geq a^2) &\leq \frac{\mathbb{E}[(X - \mu_X)^2]}{a^2} \\ &= \frac{\sigma_X^2}{a^2} \end{aligned}$$

as required.

(c) An estimator $\hat{\theta}_n$ of a parameter is said to be consistent if it converges in probability to the true parameter value θ, that is if:

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) \quad = \quad 0$$

Using the Chebyshev inequality, show (informally) that the following two conditions are sufficient to establish consistency:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_n] &= \theta \ \text{(unbiased)} \\ \lim_{n\to\infty} \text{VAR}(\hat{\theta}_n) &= 0 \end{aligned}$$

Solution: If $\hat{\theta}_n$ is unbiased, we can write

$$P(|\hat{\theta}_n - \theta| \geq \epsilon) \quad = \quad P(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| \geq \epsilon)$$

Applying the Chebyshev inequality to the RHS, we get:

$$P(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| \geq \epsilon) \quad \leq \quad \frac{\text{VAR}(\hat{\theta}_n)}{\epsilon^2}$$

From the RHS above we can see that if

$$\lim_{n\to\infty} \text{VAR}(\hat{\theta}_n) \quad = \quad 0$$

the estimator converges in probability to θ, that is, it is consistent.

(d) (*Weak Law of Large Numbers*) Let $X_1, X_2 \ldots X_n$ be a set of independently and identically distributed RVs (a *random sample*) with $\mathbb{E}[X_i] = \mu_X$ and $\text{VAR}(X_i) = \sigma_X^2 < \infty$. Use the Chebyshev inequality to show that the *sample mean*

$$\bar{X} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to the true mean μ_X.

Solution: Let $\bar{X}_n$ denote the sample mean derived from n observations. This is easily shown to be unbiased. Using the Chebyshev inequality:

$$P(|\bar{X}_n - \mu_X| \geq \epsilon) \quad \leq \quad \frac{\text{VAR}(\bar{X}_n)}{\epsilon^2}$$

But:

$$\begin{aligned} \text{VAR}(\bar{X}_n) &= \text{VAR}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) \\ &= \frac{\sigma_X^2}{n} \end{aligned}$$

Therefore

$$P(|\bar{X}_n - \mu_X| \geq \epsilon) \quad \leq \quad \frac{\sigma_X^2}{n\epsilon^2}$$

and

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu_X| \geq \epsilon) = 0$$

which means $\bar{X}_n$ converges in probability to the true mean μ_X, as required.

2. $\mathbf{X}_1 \ldots \mathbf{X}_n$, $\mathbf{X}_i \in \mathbb{R}^d$ are independently and identically distributed multivariate Normal Random Vectors, each having pdf:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

   (a) Write down the log-likelihood function for this model. (2 marks)

   (b) Derive maximum likelihood estimators for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. (4 marks)

Solution:

2(a). Log-likelihood:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad -\frac{dn}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})$$

2(b). We proceed in two stages. We first treat $\boldsymbol{\Sigma}$ as fixed, and maximize $\mathcal{L}$ to get a value $\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma})$ which maximizes $\mathcal{L}$ for a given matrix parameter $\boldsymbol{\Sigma}$. Taking the derivative of the $\mathcal{L}$ wrt vector $\boldsymbol{\mu}$, we get:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}}\mathcal{L} \quad = \quad (\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}))^T$$

Setting the derivative to zero, taking the transpose of both sides and pre-multiplying by $\boldsymbol{\Sigma}$, we get:

$$\mathbf{0} \quad = \quad \sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})$$

Solving for $\boldsymbol{\mu}$:

$$\begin{aligned} \hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \\ &= \bar{\mathbf{X}} \end{aligned}$$

Since this solution does not depend on $\boldsymbol{\Sigma}$, $\bar{\mathbf{X}}$ is the maximum likelihood estimator of $\boldsymbol{\mu}$ for any $\boldsymbol{\Sigma}$. To obtain $\hat{\boldsymbol{\Sigma}}$ we plug $\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) = \bar{\mathbf{X}}$ into the log-likelihood to obtain

$$-\frac{dn}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \tag{1}$$

and maximize this function wrt $\mathbf{\Sigma}$.

We first introduce a sample covariance matrix $\mathbf{S}$ defined as follows:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

This allows us to re-write the quadratic form in (1) as a matrix trace:

$$\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^T\mathbf{\Sigma}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) = n\,\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{S})$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of its matrix argument.

This in turn allows us to write the derivative of (1) wrt $\mathbf{\Sigma}$ as follows:

$$-\frac{n}{2}\frac{\mathrm{d}}{\mathrm{d}\mathbf{\Sigma}}\log(|\mathbf{\Sigma}|) - \frac{n}{2}\frac{\mathrm{d}}{\mathrm{d}\mathbf{\Sigma}}\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{S})$$

At this point we make use of two useful matrix derivatives (these can be found in Appendix C of Bishop and the note "Matrix Identities" by Roweis, available on the course website):

$$\begin{aligned}\frac{\partial}{\partial\mathbf{A}}\log(|\mathbf{A}|) &= (\mathbf{A}^{-1})^T \\ \frac{\partial}{\partial\mathbf{X}}\mathrm{Tr}(\mathbf{X}^{-1}\mathbf{A}) &= -\mathbf{X}^{-1}\mathbf{A}^T\mathbf{X}^{-1}\end{aligned}$$

This gives the derivative (2) in the following form (where we make use of the fact that both $\mathbf{\Sigma}^{-1}$ and $\mathbf{S}$ are symmetric):

$$-\frac{n}{2}\mathbf{\Sigma}^{-1} + \frac{n}{2}\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}$$

Setting to zero and solving, we get:

$$\begin{aligned}\hat{\mathbf{\Sigma}} &= \mathbf{S} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T\end{aligned}$$

3. $X_1 \ldots X_n, X_i \in \mathbb{R}$ are independently and identically distributed Normal RVs, each having pdf:

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

   (a) Write down the log-likelihood function for this model. (2 marks)

   (b) Derive maximum likelihood estimators for the parameters μ and σ^2. (4 marks)

Solution:

3(a). The log-likelihood function for this model is:

$$\mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

3(b). We first treat σ^2 as fixed, and maximize $\mathcal{L}$ to get a value $\hat{\mu}(\sigma^2)$ which maximizes $\mathcal{L}$ for a given value σ^2. Taking the derivative of the $\mathcal{L}$ wrt μ, setting to zero and solving, we get:

$$\begin{aligned}\hat{\mu}(\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \bar{X}\end{aligned}$$

Since this solution does not depend on σ^2, $\bar{X}$ is the maximum likelihood estimator of μ for any σ^2. We now plug this estimate into the log-likelihood and maximize the resulting function wrt σ^2 to get $\hat{\sigma}^2$. Taking the derivative wrt σ^2 and setting to zero:

$$0 = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Solving for σ^2, we get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Thus, the desired pair of MLEs is

$$\begin{aligned}(\hat{\mu}, \hat{\sigma}^2) &= \left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\ \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i\end{aligned}$$

4. Consider a classifier with Bernoulli class-conditional distributions, in which input vectors $\mathbf{X}_i \in \{0,1\}^d$ are taken to be i.i.d. given class $Y \in \{0,1\}$ and the inputs are further taken to be mutually independent (*Naive Bayes* assumption). That is, if θ_{jk} is a Bernoulli parameter giving the probability that the j^{th} input is 1, given output $Y = k$, the class-conditional distribution is:

$$P(\mathbf{X}_i \mid Y_i = k) = \prod_{j=1}^{d} \theta_{jk}^{X_{ij}}(1 - \theta_{jk})^{(1-X_{ij})}$$

(a) Write down the log-likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \log P(\mathbf{X}_1 \ldots \mathbf{X}_n \mid Y_1 \ldots Y_n, \boldsymbol{\theta})$$

for this model. Here, $\boldsymbol{\theta}$ denotes the full set of model parameters. (2 marks)

(b) Derive maximum likelihood estimates (MLEs) for the model parameters. (4 marks)

Solution:

4(a). We have n input vectors $\mathbf{X}_i$, each of dimensionality d, and conditionally independent given class labels. Let X_{ij} denote the j^{th} component of the i^{th} input vector. Considering just one input j:

$$\begin{aligned} & P(X_{1j} \ldots X_{nj} \mid Y_1 \ldots Y_n, \boldsymbol{\theta}) \\ &= \prod_{i:Y_i=1} \theta_{j1}^{X_{ij}} (1-\theta_{j1})^{(1-X_{ij})} \times \prod_{i:Y_i=0} \theta_{j0}^{X_{ij}} (1-\theta_{j0})^{(1-X_{ij})} \\ &= \theta_{j1}^{n_{j1}} (1-\theta_{j1})^{(n_1-n_{j1})} \times \theta_{j0}^{n_{j0}} (1-\theta_{j0})^{(n_0-n_{j0})} \end{aligned}$$

where, n_{jk} represents the number of observations in which the j^{th} input is 1 when the class label is k, that is:

$$n_{jk} = |\{i : X_{ij} = 1 \wedge Y_i = k\}|$$

The Naive Bayes assumption means that the probability of the complete input vector is just the product of the probabilities of the individual inputs. This means the overall conditional likelihood is just

$$\begin{aligned} & P(\mathbf{X}_1 \ldots \mathbf{X}_n \mid Y_1 \ldots Y_n, \boldsymbol{\theta}) \\ &= \prod_{j=1}^{d} \theta_{j1}^{n_{j1}} (1-\theta_{j1})^{(n_1-n_{j1})} \times \theta_{j0}^{n_{j0}} (1-\theta_{j0})^{(n_0-n_{j0})} \end{aligned}$$

and the corresponding conditional log-likelihood is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log P(\mathbf{X}_1 \ldots \mathbf{X}_n \mid Y_1 \ldots Y_n, \boldsymbol{\theta}) \\ &= \sum_{j=1}^{d} n_{j1} \log \theta_{j1} + (n_1 - n_{j1}) \log(1-\theta_{j1}) \\ &\quad + n_{j0} \log \theta_{j0} + (n_0 - n_{j0}) \log(1-\theta_{j0}) \end{aligned}$$

4(b). Taking the derivative of the log-likelihood function wrt θ_{j1} and setting to zero, we get:

$$0 = \frac{n_{j1}}{\theta_{j1}} - \frac{n_1 - n_{j1}}{1-\theta_{j1}}$$

Solving, we get the MLE $\hat{\theta}_{j1}$:

$$\hat{\theta}_{j1} = \frac{n_{j1}}{n_1}$$

Similarly:

$$\hat{\theta}_{j0} = \frac{n_{j0}}{n_0}$$