CMPS242 Homework #2 – Chapter #1 Exercises

Zayd Hammoudeh

October 10, 2017

Chapter #1, Problem #3

Suppose that we have three colored boxes, r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes; box b contains 1 apple, 1 orange, and 0 limes; box g contains 3 apples, 3 oranges, and 4 limes.

(a) **If a box is chosen at random with probabilities, $p(r) = 0.2$, $p(b) = 0.2$, and $p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple?**

Recall the Law of Total Probability establishes that a set of disjoint events, $B_1, ... B_n$, that partition the sample space, S, (i.e., $\bigcup_{(i=1)}^n B_i = S$ and $\forall_{i,j} B_i \cap B_j = \emptyset$), then for any event A in S:

$$\Pr[A] = \sum_{i=1}^{n} \Pr[B_i] * \Pr[A|B_i]$$

Hence, define B as the set of all boxes (i.e., red, blue, and green). Then, the probability of selecting an apple can be found via:

$$\Pr[apple] = \sum_{b_i \in B} \Pr[apple|b_i] * \Pr[b_i].$$

This can be rewritten as:

$$\Pr[apple] = \Pr[apple|r] * \Pr[r] + \Pr[apple|b] * \Pr[b] + \Pr[apple|g] * \Pr[g]$$
$$\Pr[apple] = 0.3 * 0.2 + 0.5 * 0.2 + 0.3 * 0.6$$
$$\Pr[apple] = \boxed{0.34}$$

(b) **If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?**

The goal of this question is to find the posterior probability, $\Pr[g|orange]$. The simplest way to do that is to use the priors, $\Pr[orange]$ and $\Pr[g]$ with the likelihood, $\Pr[orange|g]$. Hence:

$$\Pr[g|orange] = \frac{\Pr[orange|g] * \Pr[g]}{\Pr[orange]}. \tag{1}$$

Using again the Law of Total Probability, the prior probability of selecting an zzzorange, $\Pr[o]$, is:

$$\Pr[orange] = \Pr[orange|r] * \Pr[r] + \Pr[orange|b] * \Pr[b] + \Pr[orange|g] * \Pr[g]$$
$$\Pr[orange] = 0.4 * 0.2 + 0.5 * 0.2 + 0.3 * 0.6$$
$$\Pr[orange] = 0.36$$

This can then be substituted into the Eq. (1).

$$\Pr[g|orange] = \frac{\Pr[orange|g] * \Pr[g]}{\Pr[orange]}$$
$$\Pr[g|orange] = \frac{0.3 * 0.6}{0.36}$$
$$\Pr[g|orange] = \boxed{0.5}$$

Chapter #1, Problem #9

**Show that the mode (i.e., the maximum) of the Gaussian distribution (1.46) is given by μ.
Similarly, show that the mode of the multivariate Gaussian (1.52) is given by μ.**

(a)   Show the maximum of the Gaussian distribution is given by μ.

The maximizing value of a strictly positive function is equal to the maximizing value of the logarithm of the function. Below, "ln" is applied to the Gaussian formula.

$$\mathcal{N}(x|\mu,\sigma) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

$$\ln\left(\mathcal{N}(x|\mu,\sigma)\right) = \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{-(x-\mu)^2}{2\sigma^2}$$

The derivative can be applied and set to 0 as shown in Eq. (2).

$$0 = \frac{d}{dx}\ln\left(\mathcal{N}(x|\mu,\sigma)\right) = 0 - \frac{2(x-\mu)}{2\sigma^2} \tag{2}$$

Eq. (2) can be simplified yielding the final result as shown in Eq. (3).

$$x = \mu \quad \square \tag{3}$$

(b)   Show the mode of the multivariate Gaussian distribution is given by μ.

The multivariate Gaussian distribution is shown in Eq. (4). $\mathbf{x}$ is an n-dimensional input vector while $\boldsymbol{\mu}$ is the n-dimensional mean vector. Σ is the $n \times n$ covariance matrix.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) := \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{4}$$

Similar to the technique used in part (a), the maximizing (mode) value of a strictly positive function is also the maximizing value of the function's natural log. Hence, Eq. (4) becomes:

$$\ln\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right) \tag{5}$$

$$= \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) + \ln\left(\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right) \tag{6}$$

$$= \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{7}$$

The gradient of an n-dimensional function, $f(\mathbf{x})$, is an n-dimensional vector as defined in Eq. (8).

$$\nabla f(\mathbf{x}) = \left\langle \frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n} \right\rangle \tag{8}$$

A multivariate function is maximized when its gradient equals the zero vector (i.e., $\mathbf{0}$). Hence, take the derivative of Eq. (7) and set it equal to the zero vector, which yields Eq. (10).

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{x}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \mathbf{0} + \frac{\partial}{\partial \mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{9}$$

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{10}$$

If $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$ and $A = \Sigma^{-1}$, then Eq. (10) becomes:

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{y}} \mathbf{y}^{\mathrm{T}} A \mathbf{y} \tag{11}$$

This matches the well known derivative identity that:

$$\frac{\partial}{\partial \mathbf{y}} \mathbf{y}^{\mathrm{T}} A \mathbf{y} = (A^{\mathrm{T}} + A) y.$$

Hence, Eq. (11) simplifies to:

$$\mathbf{0} = (A^{\mathrm{T}} + A) \mathbf{y}.$$

Substituting back in for $\mathbf{y}$ completes the proof via:

$$(A^{\mathrm{T}} + A)(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$$

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{0}$$

$$\mathbf{x} = \boldsymbol{\mu}. \quad \square$$

Chapter #1, Problem #11

By setting the derivatives of the log likelihood function (1.54) with respect to μ and σ^2 equal to zero, verify the results (1.55) and (1.56).

The log likelihood function (Equation 1.54) is defined on page 27 as:

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{12}$$

(a) **Verify Equation (1.55) that:**

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

Take the partial derivative with respect to μ:

$$\frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) = \frac{-1}{2\sigma^2}\left(2*-1*\sum_{n=1}^{N}(x_n - \mu)\right) - 0 - 0$$

$$= \frac{\sum_{n=1}^{N}(x_n - \mu)}{\sigma^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n)}{\sigma^2} - \frac{N*\mu}{\sigma^2}.$$

The right side of the equation can be set equal to 0. The denominator can be multiplied out resulting in:

$$0 = \sum_{n=1}^{N}(x_n) - N*\mu$$

$$N*\mu = \sum_{n=1}^{N}x_n$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N}x_n \quad \square$$

(b) **Verify Equation (1.56) that:**

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Taking the partial derivative with respect to σ and using the chain rule as necessary, we get:

$$\frac{\partial}{\partial\sigma}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) = -\frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 * (-2*\sigma^{-3}) - \frac{N}{2}\left(\frac{1}{\sigma^2}(2\sigma)\right) - 0$$

$$= \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{\sigma^3} - \frac{N}{\sigma}$$

This equation is set equal to 0 and everything multiplied by σ resulting in:

$$0 = \frac{\sum_{n=1}^{N} (x_n - \mu)^2}{\sigma^2} - N$$

$$N = \frac{\sum_{n=1}^{N} (x_n - \mu)^2}{\sigma^2}$$

$$\sigma^2 = \frac{\sum_{n=1}^{N} (x_n - \mu)^2}{N}.$$

The optimal value for μ was found in part (a) which can then be substituted into the previous equation yielding:

$$\sigma^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_{ML})^2}{N}. \quad \square$$

Chapter #1, Problem #29

Consider an M-state discrete random variable, x, and use Jensen's inequality in the form (1.115) to show that the entropy of its distribution $p(x)$ satisfies $\mathbf{H}[x] \leq \ln M$.

Equation (1.115) in the textbook (page 56) states:

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i).$$

A probability distribution, $p(x)$, satisfies the two Jensen Inequality criteria of λ namely: $\forall_i \lambda_i \geq 0$ and $\sum_{i=1}^{M} (\lambda_i) = 1$. This transforms Jensen's Inequality to:

$$f\left(\sum_{i=1}^{M} p(x_i) x_i\right) \leq \sum_{i=1}^{M} p(x_i) f(x_i). \tag{13}$$

Recall that a random variable is a function, $X : \Omega \to \mathbb{R}$, where Ω is the sample space (in this case $|\Omega| = M$) and $\mathbb{R}$ is the set of real numbers. For this proof, define

$$x_i = \frac{1}{p(x_i)}.$$

This can be substituted into Eq. (13) yielding:

$$f\left(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)}\right) \leq \sum_{i=1}^{M} p(x_i) f\left(\frac{1}{p(x_i)}\right)$$

$$f\left(\sum_{i=1}^{M} 1\right) \leq \sum_{i=1}^{M} p(x_i) f\left(\frac{1}{p(x_i)}\right)$$

$$f(M) \leq \sum_{i=1}^{M} p(x_i) f\left(\frac{1}{p(x_i)}\right).$$

Define $f(x) = \ln\left(\frac{1}{x}\right)$ yielding:

$$\ln\left(\frac{1}{M}\right) \leq \sum_{i=1}^{M} p(x_i) \ln p(x_i).$$

Multiplying both sides by -1 (which inverts the inequality) and using the definition of entropy $H[x] = -\sum p(x) \ln p(x)$ completes the proof.

$$-\ln\left(\frac{1}{M}\right) \geq -\sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}$$

$$\ln M \geq \mathbf{H}[x]. \quad \square$$

Chapter #1, Problem #30
Evaluate the Kullback-Leibler divergence (1.113) between two Gaussians, $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, s^2)$.

For two continuous probability distributions, p and q, the Kullback-Leibler divergence is defined as:

$$\mathrm{KL}(p||q) := \int p(x) \ln \frac{p(x)}{q(x)} dx.$$

Let's simplify the logarithm first. Substituting for $p(x)$ and $q(x)$ yields:

$$\ln \left(\frac{p(x)}{q(x)} \right) = \ln \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)}{\frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-m)^2}{2s^2} \right)} \right)$$

$$= \ln \left(\frac{s}{\sigma} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-m)^2}{2s^2} \right) \right)$$

$$= \ln \left(\frac{s}{\sigma} \right) + \ln \left(\exp \left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-m)^2}{2s^2} \right) \right)$$

$$= \ln \left(\frac{s}{\sigma} \right) - \frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-m)^2}{2s^2}.$$

Substituting this back into the original equation yields:

$$\mathrm{KL}(p||q) = \int p(x) \left(\ln \left(\frac{s}{\sigma} \right) - \frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-m)^2}{2s^2} \right) dx.$$

Expand the equation for simplification.

$$\mathrm{KL}(p||q) = \int p(x) \ln \left(\frac{s}{\sigma} \right) - p(x) \frac{(x^2 - 2 \cdot x \cdot \mu + \mu^2)}{2\sigma^2} + p(x) \frac{(x^2 - 2 \cdot x \cdot m + m^2)}{2s^2} dx. \tag{14}$$

Given $p(x) = \mathcal{N}(x|\mu, \sigma^2)$, a few useful identities are in Eq. (15), (16), and (17).

$$\int c \cdot p(x) dx = 1 \tag{15}$$

$$\int x \cdot p(x) dx = \mu \tag{16}$$

$$\int x^2 \cdot p(x) dx = \mu^2 + \sigma^2 \tag{17}$$

Using these identities, Eq. (14) simplifies to:

$$\mathrm{KL}(p||q) = \ln \left(\frac{s}{\sigma} \right) - \frac{(\mu^2 + \sigma^2 - 2 \cdot \mu^2 + \mu^2)}{2\sigma^2} + \frac{(\mu^2 + \sigma^2 - 2 \cdot \mu \cdot m + m^2)}{2s^2}$$

$$= \ln \left(\frac{s}{\sigma} \right) - \frac{1}{2} + \frac{\sigma^2 + (\mu - m)^2}{2s^2}$$

If desired, the fraction can be pulled out resulting in the final form in Eq. (18).

$$\mathrm{KL}(p||q) = \frac{1}{2} \left(\ln \left(\frac{s^2}{\sigma^2} \right) - 1 + \frac{\sigma^2 + (\mu - m)^2}{s^2} \right) \tag{18}$$

Chapter #1, Problem #31

Consider two variables x and y having joint distribution $p(\mathbf{x}, \mathbf{y})$, show that the differential entropy of this pair of variables satisfies

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] \leq \mathbf{H}[\mathbf{x}] + \mathbf{H}[\mathbf{y}]$$

with equality if and only if x and y are statistically independent.

The joint entropy of two variables can be written in terms of their marginal entropies of and their mutual information, $\mathbf{I}[\mathbf{x}, \mathbf{y}]$. This is shown in Eq. (19).

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] + H[\mathbf{y}] - I[\mathbf{x}; \mathbf{y}]. \tag{19}$$

Cover & Thomas show that mutual information has the relationship:

$$\mathbf{I}[\mathbf{x}, \mathbf{y}] \geq 0 \tag{20}$$

with equality only in the case that $\mathbf{x}$ and $\mathbf{y}$ are independent (see (2.90) on page 28 of the second edition of Cover & Thomas). Given Eq. (20) and mutual information's non-negativity, it is clear that:

$$H[\mathbf{x}] + H[\mathbf{y}] - I[\mathbf{x}; \mathbf{y}] \leq H[\mathbf{x}] + H[\mathbf{y}].$$

This in turn can be combined with Eq. (19) yielding:

$$H[\mathbf{x}, \mathbf{y}] \leq H[\mathbf{x}] + H[\mathbf{y}]. \tag{21}$$

Equality holds when $\mathbf{I}[\mathbf{x}; \mathbf{y}] = 0$, which only occurs when $\mathbf{X}$ and $\mathbf{Y}$ are independent per Cover & Thomas. $\square$

Chapter #1, Problem #40

By applying Jensen's inequality (1.115) with $f(x) = \ln x$, show that the arithmetic mean of a set of real numbers is never less than their geometrical mean.

For a set of numbers, $\{x_1, ..., x_M\}$, the arithmetic mean is defined as:

$$\frac{\sum_{i=1}^{M} x_i}{M}.$$

In contrast, for the same set of numbers, the geometric mean is defined as:

$$\left(\prod_{i=1}^{M} x_i\right)^{\frac{1}{M}}.$$

Jensen's Inequality (1.115) is defined as:

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i).$$

If we substitute for $f(x) = \ln x$, the inequality becomes:

$$\ln\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i \ln(x_i).$$

For Jensen's Inequality to hold, λ must satisfy two conditions namely: $\lambda_i \geq 0$ and $\sum_{i=1}^{M} (\lambda_i) = 1$. An obvious satisfying case is $\lambda_i = 1/M$ for all $i = 1...M$. This then changes the equation to:

$$\ln\left(\frac{\sum_{i=1}^{M} x_i}{M}\right) \leq \frac{1}{M} \sum_{i=1}^{M} \ln(x_i).$$

Using the properties of logarithms, the right side is transformable to a product via:

$$\ln\left(\frac{\sum_{i=1}^{M} x_i}{M}\right) \leq \frac{1}{M} \ln\left(\prod_{i=1}^{M} x_i\right).$$

Using another property of logarithms, the multiplying scalar, $\frac{1}{M}$ can be brought inside the logarithm as:

$$\ln\left(\frac{\sum_{i=1}^{M} x_i}{M}\right) \leq \ln\left(\left(\prod_{i=1}^{M} x_i\right)^{\frac{1}{M}}\right).$$

Both sides are then raised to the power of e completing the proof.

$$\frac{\sum_{i=1}^{M} x_i}{M} \leq \left(\prod_{i=1}^{M} x_i\right)^{\frac{1}{M}} \quad \square$$

Extra Problem #1
Prove that the sum of two convex functions is also convex.

A function, $f(x)$, is convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{22}$$

Given a second convex function, $g(x)$, and an additional function, $h(x) = f(x) + g(x)$, add the definition of convexity for f to the definition of convexity for g as shown in Eq. (23).

$$f(\lambda x_1 + (1 - \lambda)x_2) + g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) + \lambda g(x_1) + (1 - \lambda)g(x_2). \tag{23}$$

This can be rearranged due to the associativity and commutativity of real valued addition.

$$(f(\lambda x_1 + (1 - \lambda)x_2) + g(\lambda x_1 + (1 - \lambda)x_2)) \leq \lambda(f(x_1)) + g(x_1)) + (1 - \lambda)(f(x_2) + g(x_2)).$$

Substituting using the definition of $h(x)$ proves convexity via:

$$h(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda h(x_1) + (1 - \lambda)h(x_2). \quad \square$$