Assignment One

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

October 8, 2017

1 Probability and Independence

For this question we prove or disprove the following properties of independence.

1.1 $(X \perp Y, W \mid Z)$ IMPLIES $(X \perp Y \mid Z)$ is TRUE

PROOF: Using the law of total probability and the conditional independence of X and the joint Y, W on Z it is clear that:

$$\begin{aligned} P(X, Y \mid Z) = \int_W P(X, Y, W \mid Z) dW &= \int_W P(X \mid Z) P(Y, W \mid Z) dW \\ &= P(X \mid Z) \int_W P(Y, W \mid Z) dW \\ &= P(X \mid Z) P(Y \mid Z) \end{aligned}$$

1.2 $(X \perp Y \mid Z)$ AND $(X, Y \perp W \mid Z)$ IMPLIES $(X \perp W \mid Z)$ is TRUE

PROOF: Notice that, by symmetry, $(X, Y \perp W \mid Z) \Rightarrow (W \perp X, Y \mid Z)$, and so by 1.1 we have that $(W \perp X \mid Z)$ (and hence $(X \perp W \mid Z)$)

1.3 $(X \perp Y, W \mid Z)$ AND $(Y \perp W \mid Z)$ IMPLIES $(X, W \perp Y \mid Z)$ is TRUE

PROOF: First, notice $(X \perp Y, W \mid Z) \implies (X \perp W \mid Z)$ by 1.1. Then, clearly:

$$\begin{aligned} P(X, Y, W \mid Z) = P(X \mid Z)P(Y, W \mid Z) &= P(X \mid Z)P(Y \mid Z)P(W \mid Z) \\ &= P(X, W \mid Z)P(Y \mid Z) \end{aligned}$$

1.4 $(X \perp Y \mid Z)$ AND $(X \perp Y \mid W)$ IMPLIES $(X \perp Y \mid W, Z)$ is FALSE

COUNTER EXAMPLE: Take the following Probability space: $\Omega = \{1, 2, 3, \cdots, 16\}$ as the Sample space and 2^Ω as the σ-algebra equipped with the Counting Measure. Consider the following random variables:

$$\begin{aligned} Z &= \mathbb{1}_{\{1,2,3,4,7,8,11,12,13\}} \\ W &= \mathbb{1}_{\{1,2,5,6,9,10,14,15,16\}} \\ X &= \mathbb{1}_{\{1,3,5,7,9\}} \\ Y &= \mathbb{1}_{\{2,7,8,9,10\}} \end{aligned}$$

Notice that

$$\begin{aligned} P(X, Y \mid Z) &= \frac{1}{9} = P(X \mid Z)P(Y \mid Z) \\ P(X, Y \mid W) &= \frac{1}{9} = P(X \mid W)P(Y \mid W) \end{aligned}$$

but

$$P(X, Y \mid Z, W) = 0 \neq P(X \mid Z, W)P(Y \mid Z, W) = \frac{1}{4}$$

2 Bayesian Inference and MAP

2.1 Given IID Data, what are the conditional independence properties for $P(\pi, x_1, \cdots, x_n)$?

Denote $X_A := \{X_i\}_{i \in A}$. We can see that $(X_A \perp X_B \mid \pi)$ for any non-intersecting collection $A, B \subset \{1, 2, \cdots, n\}$, since, using the IID property

$$\begin{aligned} P(X_A, X_B \mid \pi) &= \prod_{i \in A \sqcup B} P(X_i \mid \pi) \\ &= P(X_A \mid \pi)P(X_B \mid \pi) \end{aligned}$$

i.e. the X_i's are mutually independent, conditioned on π.

2.2 Derive $P(\pi \mid x^{(1)}, \cdots, x^{(n)})$

Note that, ignoring normalization constants we can see that:

$$\begin{aligned} P(\pi \mid x^{(1)}, \cdots, x^{(n)}) &\propto P(x^{(1)}, \cdots, x^{(n)} \mid \pi) P(\pi \mid \alpha) \\ &\propto \prod_{j=1}^{k} \pi_j^{\sum_{i=1}^{n} x_j^{(i)}} \prod_{j=1}^{k} \pi_j^{\alpha_j - 1} = \prod_{j=1}^{k} \pi_j^{N_j + \alpha_j - 1} \end{aligned}$$

Where $N_j = \sum_{i=1}^{n} x_j^{(i)}$. We recognize the above as the unnormalized density of a Dirichlet Random Variable, and so $\pi \mid x^{(1)}, \cdots, x^{(n)} \sim Dirichlet(\{N_j + \alpha_j\}_{j=1}^{k})$

2.3 Derive $P(x^{(1)}, \cdots, x^{(n)})$

To compute the data density we integrate out π

$$\begin{aligned} P(x^{(1)}, \cdots, x^{(n)}) &= \int_{\pi \in \Delta_k} P(\pi, x^{(1)}, \cdots, x^{(n)}) d\pi \\ &= \int_{\pi \in \Delta_k} P(x^{(1)}, \cdots, x^{(n)} \mid \pi) P(\pi \mid \alpha) d\pi \\ &= \int_{\pi \in \Delta_k} \prod_{j=1}^{k} \pi_j^{N_j} \frac{1}{C(\alpha)} \prod_{j=1}^{k} \pi_j^{\alpha_j - 1} \end{aligned}$$

Where $C(\alpha) = C(\alpha_1, \cdots, \alpha_k) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j)}$

$$\begin{aligned} &= \frac{1}{C(\alpha)} \int_{\pi \in \Delta_k} \prod_{j=1}^{k} \pi_j^{N_j + \alpha_j - 1} \\ &= \frac{C(N + \alpha)}{C(\alpha)} \end{aligned}$$

Since $\frac{1}{C(N+\alpha)} \prod_{j=1}^{k} \pi_j^{N_j + \alpha_j - 1} \sim Dirichlet(N + \alpha)$

2.4 Derive MAP estimate $\hat{\pi}$ for π and compare it to the MLE estimator

We assume that $\alpha_j > 1 \ \forall j$.
Notice that

$$\begin{aligned} \max_{\pi \in \Delta_k} P(\pi \mid x^{(1)}, \cdots, x^{(n)}) &= \max_{\pi \in \Delta_k} P(x^{(1)}, \cdots, x^{(n)} \mid \pi) P(\pi \mid \alpha) \\ &= \max_{\pi \in \Delta_k} \prod_{j=1}^{k} \pi_j^{N_j + \alpha_j - 1} \end{aligned}$$

And using the monotonicity of log along with the definition of Δ_k we can write this as the equivalent constrained optimization:

$$\begin{gathered}\text{Maximize } l(\pi) = \sum_{j=1}^{k}(N_j + \alpha_j - 1)log(\pi_j) \\ \text{subject to the constraint } g(\pi) = 1 - \sum_{i=1}^{k}\pi_i = 0\end{gathered}$$

We solve this using the method of Lagrange Multipliers. We look for any stationary points i.e.

$$\begin{aligned}\nabla_\pi(l(\pi) + \lambda g(\pi)) = 0 \\ \nabla_\lambda(l(\pi) + \lambda g(\pi)) = 0\end{aligned}$$

And solving for each π_j yields

$$\begin{aligned}\nabla_{\pi_j}(l(\pi_j) + \lambda g(\pi_j)) = \frac{N_j + \alpha_j - 1}{\pi_j} - \lambda = 0 \\ \Rightarrow \pi_j = \frac{N_j + \alpha_j - 1}{\lambda}\end{aligned}$$

And solving for λ yields

$$\begin{aligned}\nabla_\lambda(l(\pi) + \lambda g(\pi)) = 1 - \sum_{i=1}^{k}\pi_i = 0 \\ \Rightarrow 1 - \sum_{i=1}^{k}\frac{N_i + \alpha_i - 1}{\lambda} = 0 \\ \Rightarrow \lambda = \sum_{i=1}^{k} N_i + \alpha_i - 1 = N + \sum_{i=1}^{k}(\alpha_i - 1)\end{aligned}$$

And putting the above together, we get that each $\pi_j = \frac{N_j+\alpha_j-1}{N+\sum_{i=1}^{k}(\alpha_i-1)}$.
To check that our stationary point is a maximum, we compute the determinent of the Hessian and check if it is negative.

If we compared $\hat{\pi}^{MLE}$ with $\hat{\pi}^{MAP}$, as k becomes very large, for any k, $\hat{\pi_k}^{MAP}$ will shrink since the probabilities of each $\hat{\pi_k}^{MAP}$ will be necessarily non-zero, whereas $\hat{\pi_j}^{MLE} = 0$ whenever $N_k = 0$ (which would be very frequent when k gets larger than N).

3 Properties of estimators

3.1 Find MLE of $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Poisson(\lambda)$ and determine its Bias, Variance and Consistency

We want to find a λ to maximize the following:

$$L(\lambda \mid x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

to simplify the above, we take the log:

$$l(\lambda) = -n\lambda + log(\lambda)\sum_{i=1}^{n} x_i - log(x_i!)$$

We differentiate with respect to λ and find the stationary point λ^*.

$$\begin{aligned} \partial l(\lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\ \Rightarrow \lambda^* = \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

The second derivative is negative for all $\lambda \neq 0$, and so λ^* is a maximum.

$$\partial^2 l(\lambda) = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$$

We compute the mean and variance of λ^{MLE}

$$E(\lambda^{MLE}) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\lambda}{n} = \lambda \tag{3.1}$$

$$Var(\lambda^{MLE}) = \frac{Var(\sum_{i=1}^{n} X_i)}{n^2} = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} = \frac{\lambda}{n} \tag{3.2}$$

From (3.1) we see that λ^{MLE} is unbiased, and using (3.2) and $Var(\lambda^{MLE}) \to 0$, we have that λ^{MLE} is consistent.

3.2 Given $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Bern(p)$ and estimator $\hat{p} := \frac{1}{10}\sum_{i=1}^{10} X_i$ determine its Bias, Variance and Consistency

We compute the mean and variance of $\hat{p}$

$$E(\hat{p}) = \frac{1}{10}\sum_{i=1}^{10} E(X_i) = \frac{10p}{10} = p \tag{3.3}$$

$$Var(\hat{p}) = \frac{1}{100} Var\left(\sum_{i=1}^{10} X_i\right) = \frac{p(1-p)}{10} \tag{3.4}$$

And we use $\sum_{i=1}^{10} X_i \sim Bin(10p, 10p(1-p))$ for (3.4). We can see that $\hat{p}$ is unbiased, but is not consistent in l^2, since $E(||\hat{p} - p||^2) = Bias(\hat{p})^2 + Var(\hat{p}) = \frac{p(1-p)}{10} \not\to 0$

3.3 Given $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Unif(\theta)$ find MLE and determine its Bias, Variance and Consistency

We want to find a θ to maximize the following:

$$\begin{aligned} L(\theta \mid x_1, x_2, \cdots, x_n) &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{\{x_i \leq \theta\}} \\ &= \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} \mathbb{1}_{\{x_i \leq \theta\}} \\ &= \left(\frac{1}{\theta}\right)^n \mathbb{1}_{\{x_1 \leq \theta, \cdots, x_n \leq \theta\}} \\ &= \left(\frac{1}{\theta}\right)^n \mathbb{1}_{\{x_{(n)} \leq \theta\}} \end{aligned}$$

where $x_{(n)}$ is the nth order statistic.
Notice that as θ decreases $L(\theta \mid x_1, x_2, \cdots, x_n)$ increases, but we cannot take θ^{MLE} to be arbitrarily small, since $L(\theta \mid x_1, x_2, \cdots, x_n) = 0$ whenever $\theta < x_{(n)}$. Hence $\theta^{MLE} = x_{(n)}$.

Note that $F_{x_{(n)}}(x) = P(x_{(n)} < x) = P(x_1 < x) \cdots P(x_n < x) = \left(\frac{x}{\theta}\right)^n \mathbb{1}_{\{x \leq \theta\}}$

And so

$$f_{x_{(n)}}(x) = \partial F_{x_{(n)}}(x) = \begin{cases} \frac{n}{\theta^n} x^{n-1} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases}$$

We compute the mean and variance of θ^{MLE}

$$\begin{aligned} E(\theta^{MLE}) &= \int_0^\theta \frac{n}{\theta^n} x^n dx = \frac{n}{n+1} \frac{x^{n+1}}{\theta^n} \Big|_0^\theta = \frac{n}{n+1}\theta \\ E((\theta^{MLE})^2) &= \int_0^\theta \frac{n}{\theta^n} x^{n+1} dx = \frac{n}{n+2} \frac{x^{n+2}}{\theta^n} \Big|_0^\theta = \frac{n}{n+2}\theta^2 \\ Var(\theta^{MLE}) &= E((\theta^{MLE})^2) - E(\theta^{MLE})^2 \\ &= \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2} \end{aligned}$$

So θ^{MLE} is biased. It's consistent, though, since $E(\theta^{MLE}) \to \theta$ and $Var(\theta^{MLE}) \to 0$

3.4 Given $X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ show that the MLE $(\hat{\mu}, \hat{\sigma}^2)$ is $\left(\bar{X}, \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}\right)$ and determine the Bias, Variance and Consistency for $\hat{\sigma}^2$

It is enough to show that $(\hat{\mu}, \hat{\sigma}^2)$ satisfy $\nabla l(\hat{\mu}, \hat{\sigma}^2) = 0$ and $\nabla^2 l(\hat{\mu}, \hat{\sigma}^2) = 0$ is a negative definite matrix.

This can be seen by computing the partial derivatives of
$l(\mu, \sigma^2) = \frac{n}{2} log(2\pi) - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \tag{3.5}$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.6}$$

$$\frac{\partial^2 l(\mu, \sigma^2)}{\partial \mu^2} = -\frac{n}{\sigma^2} \tag{3.7}$$

$$\frac{\partial^2 l(\mu, \sigma^2)}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.8}$$

$$\frac{\partial^2 l(\mu, \sigma^2)}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \tag{3.9}$$

Assuming $\sigma^2 \neq 0$. After setting (3.5) to 0 we get

$$\sum_{i=1}^{n} x_i = n\mu \Rightarrow \mu = \bar{X} \tag{3.10}$$

Setting (3.6) to 0 yields

$$-\frac{n}{2(\sigma^2)^2}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{3.11}$$

And substituting μ with $\bar{X}$ from (3.11) gives us

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{3.12}$$

Hence $\nabla l(\hat{\mu}, \hat{\sigma}^2) = 0$. Finally to show that $\nabla^2 l(\hat{\mu}, \hat{\sigma}^2)$ is negative definite, we substitute $\hat{\mu}, \hat{\sigma}^2$ from (3.10) and (3.12) into (3.7), (3.8), and (3.9) to obtain the following:

$$\nabla^2 l(\hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^2} \end{bmatrix}$$

It is clear that this is negative definite, since its eigenvalues are both negative. Next we find the mean and variance of the MLE $\hat{\sigma}^2$.

$$E(\hat{\sigma}^2) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right) \tag{3.13}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}E(X_i^2) - nE(\bar{X}^2)\right) = E(X_i^2) - E(\bar{X}^2) \tag{3.14}$$

$$= \mu^2 + \sigma^2 - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2 \tag{3.15}$$

$$Var(\hat{\sigma}^2) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{(\sigma^2)^2}{n^2}Var\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}\right) \tag{3.16}$$

$$= \frac{(\sigma^2)^2}{n^2}2(n-1) \tag{3.17}$$

Where (3.15) comes from the Central Limit Theorem, i.e. that $\bar{X} \sim N(\mu, \frac{\sigma}{n})$ and (3.17) comes from $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.

We can see from (3.15) that $\hat{\sigma}^2$ is biased. It's consistent, however, since $E(\hat{\sigma}^2) \to \sigma^2$ and $Var(\hat{\sigma}^2) \to 0$.

4 Empirical Experimentation

We simulated 10,000 observations from a standard Gaussian distribution and to assess the theoretical results for $\hat{\sigma}^2$ as computed in section 3.4.
First, we note that the empirical distribution plotted in figure 4.1 looks χ^2_4 distributed, as it should, given the claim used in (3.17).
Secondly we compute the bias and variance of $\hat{\sigma}^2$.
Using (3.15) we compute the Bias as

$$E(\hat{\sigma}^2 - \sigma^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} = -\frac{1}{5}$$

Our estimate of the bias was -0.20137 – which is quite close.
Using (3.17) we compute the variance as

$$Var(\hat{\sigma}^2) = \frac{(\sigma^2)^2}{n^2}2(n-1) = \frac{8}{25} = 0.32$$

Our estimate of the variance was 0.3188. Again, this is quite close.

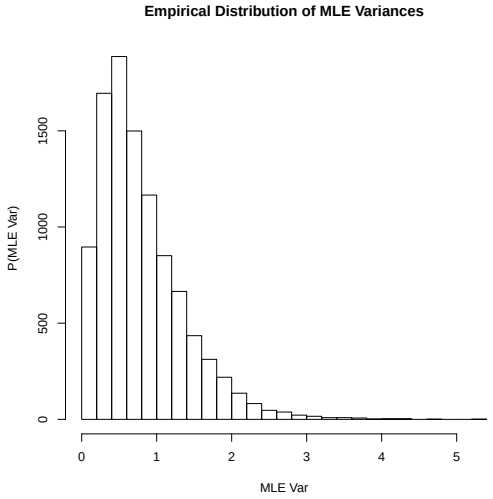

Figure 4.1: Emiprical Distribution of $\hat{\sigma}^2$