EE5907/EE5027 Week 2: Probability Review

The following questions are from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

Exercise 2.6: Conditional independence

(a) Let $H \in \{1, \cdots, K\}$ be a discrete random variable, amd let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), \cdots, P(H = K|e_1, e_2)) \tag{1}$$

Which of the following sets of numbers are sufficient for the calculation?

i. $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$ $\quad P(H|e_1,e_2) = \dfrac{P(e_1,e_2|H)\cdot P(H)}{P(e_1,e_2)}$

ii. $P(e_1, e_2), P(H), P(e_1, e_2|H)$

iii. $P(e_1|H), P(e_2|H), P(H)$ $\quad$ (ii)

$P(e_1,e_2|H) = P(e_1|H)\cdot P(e_2|H)$

(b) Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficent now? $\quad$ (i)

Show your claculations as well as giving the final result. Hint: use Bayes rule.

$P(e_1,e_2) = \sum P(e_1,e_2|H)\cdot P(H)$

Exercise 2.7: Pairwise independence does not imply mutual independence

We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{3}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \ \forall S \subseteq \{1, \cdots, n\} \setminus \{i\} \tag{4}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{5}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

Exercise 2.8: Conditional indepence iff joint factorizes

In the text we said $X \perp Y|Z$ iff

$$p(x, y|z) = p(x|z)p(y|z) \tag{6}$$

for all x, y, z such that $p(z) > 0$. Now prove the following alternative definition: $X \perp Y|Z$ iff there exist function g and h such that

$$p(x, y|z) = g(x, z)h(y, z) \tag{7}$$

for all x, y, z such that $p(z) > 0$

$\Rightarrow$ $p(x,y|z) = p(x|z) \cdot p(y|z)$. Let $g(x,z) = p(x|z)$, $h(y,z) = p(y|z)$

$\Leftarrow$ Integrade both side of x

$$\int p(x,y|z)\,dx = \int g(x,z)\,dx \cdot h(y,z)$$

$$p(y|z) = G(z) \cdot h(y,z) \qquad \text{E.q. 1}$$

Integrade both side of y

$$\int p(x,y|z)\,dy = g(x,z) \cdot \int h(y,z)\,dy$$

$$p(x|z) = g(x,z) \cdot H(z) \qquad \text{E.q 2}$$

Full integrade

$$\iint p(x,y|z)\,dx\,dy = \iint g(x,z)\,dx\, h(y,z)\,dy$$

$$1 = G(z) \cdot H(z) \qquad \text{E.q 3.}$$

$$p(x,y|z) = \frac{p(y|z)}{G(z)} \cdot \frac{p(x|z)}{H(z)} = p(y|z) \cdot p(x|z)$$

EE5907/EE5027 Week 2: MLE + MAP

The following questions are from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

Exercise 3.1 MLE for the Bernoulli/ binomial model

Derive

$$\hat{\theta}_{MLE} = \frac{N1}{N} \tag{1}$$

by optimizing the log of the likelihood in Eq. (2)

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{2}$$

$\log P(D|\theta) = N_1 \log\theta + N_0 \log(1-\theta)$

$\underset{\{\theta\}}{\text{maxarg}}\, P(D|\theta) = \underset{\{\theta\}}{\text{argmax}} \log P(D|\theta)$

$= \underset{\{\theta\}}{\text{argmax}}\{N_1 \log\theta_1 + N_0 \log(1-\theta)$

$= \underset{\{\theta\}}{\text{argmin}}[-N_1 \log\theta_1 - N_0 \log(1-\theta))$

differentiating with θ, set to 0

$\frac{N_1}{\theta} - \frac{N_0}{1-\theta} \Rightarrow$

$N_1 - N_1\theta = N_0\theta$

$\theta = \frac{N_1}{N_0+N_1} = \frac{N_1}{N}$

Exercise 3.6 MLE for the Poisson distribution

The Poisson pmf is defined as $\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \cdots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

Exercise 3.7 Bayesian analysis of the Poisson distribution

In the previous exercise, we defined the Poisson distribution with rate λ and derived its MLE. Here we perform a conjugate Bayesian analysis.

a. Derive the posterior $p(\lambda|\mathcal{D})$ assuming a conjugate prior $p(\lambda) = Ga(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b}$. Hint: the posterior is also a Gamma distribution.

b. What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $\text{Ga}(a, b)$ distribution is a/b.)

Exercise 3.6 MLE for the Poisson distribution

The Poisson pmf is defined as $\mathrm{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \cdots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

$D = x \in \{0, 1, 2 \ldots\}$

$$P(D|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$$

$$\log P(D|\lambda) = \sum(-\lambda + x_i \log\lambda - \log x_i!)$$
$$= -n\lambda + \log\lambda \cdot \sum x_i - \sum \log x_i!$$

$$\hat{\lambda}_{MLE} = \underset{\{\lambda\}}{\mathrm{argmax}}\{-n\lambda + \log\lambda \sum x_i - \sum \log x_i!\}$$
$$= \mathrm{argmax}\{-n\lambda + \log\lambda \cdot \sum x_i\}$$

differentiating with respect to λ and set to 0.

$$-n + \frac{1}{\lambda} \cdot \sum x_i = 0$$
$$\lambda = \frac{1}{n} \cdot \sum x_i = E[x_i]$$

Exercise 3.7 Bayesian analysis of the Poisson distribution

In the previous exercise, we defined the Poisson distribution with rate λ and derived its MLE. Here we perform a conjugate Bayesian analysis.

a. Derive the posterior $p(\lambda|\mathcal{D})$ assuming a conjugate prior $p(\lambda) = Ga(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b}$. Hint: the posterior is also a Gamma distribution.

b. What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $\mathrm{Ga}(a, b)$ distribution is a/b.)

$$P(\lambda|D) \propto P(D|\lambda) \cdot P(\lambda)$$
$$\propto \prod e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \cdot \lambda^{a-1} e^{-\lambda b}$$
$$\propto \prod \frac{\lambda^{x_i}}{x_i!} \cdot \lambda^{a-1} \cdot e^{-\lambda(b+n)}$$
$$\propto \prod \frac{1}{x_i!} \cdot \lambda^{\sum x_i + a-1} \cdot e^{-\lambda(b+n)}$$
$$\propto \lambda^{\sum x_i + (a-1)} \cdot e^{-\lambda(b+n)}$$

$$P(\lambda|D) = Ga(\lambda \,|\, \sum x_i + a, b+n)$$

b. $P(\lambda|D) = \dfrac{P(D|\lambda)\cdot P(\lambda)}{P(D)}$

$$\hat{\lambda}_{MLE} = \underset{\{\lambda\}}{\operatorname{argmax}}\ P(D|\lambda)\cdot P(\lambda)$$

$$= \underset{\{\lambda\}}{\operatorname{argmax}}\ P(D|\lambda)\cdot \lambda^{a-1}\cdot e^{-\lambda b}$$

$$= \underset{\{\lambda\}}{\operatorname{argmax}}\ \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\quad \lambda^{a-1}\cdot e^{-\lambda b}$$

$$= \underset{\{\lambda\}}{\operatorname{argmax}}\ \prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}\ \cdot\ \lambda^{a-1}\cdot e^{-\lambda(b+n)}$$

$$= \underset{\{\lambda\}}{\operatorname{argmax}}\left\{(a-1)\cdot\log\lambda - \lambda(b+n) + \sum x_i\log\lambda - \sum\log x_i!\right\}$$

$$\frac{a-1}{\lambda} - (b+n) + \frac{1}{\lambda}\cdot\sum x_i = 0$$

$$\lambda(b+n) = (a-1) + \sum x_i \Rightarrow \lambda = \frac{a-1+\sum x_i}{b+n}$$

Exercise 3.12 MAP estimation for the Bernouli with non-conjugate Priors

We discussed Bayesian inference of a Bernoulli rate parameter with the prior $p(\theta) = Beta(\theta|\alpha, \beta)$. We know that, with this prior, the MAP estimate is given by

$$\hat{\theta} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} \tag{3}$$

where N_1 is the number of heads, N_0 is the number of tails, and $N = N_0 + N_1$ is the total number of trials.

Now consider the following prior, that believes the coin is fair, or is slightly biased towards tails:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Derive the MAP estimate under the prior as a function of N_1 and N.

$0.4^{N_1} \cdot 0.6^{N_0}$