Machine Learning with Networks ECEN765.600, Fall 2017

Homework (Written) #1: Math Refresher & Bayesian Learning

"On my honor, as an Aggie, I have neither given nor received unauthorized aid on this academic work."

Signature: Joseph Riad

Name: Joseph Riad

Problem 1

Define the following events:

$$\begin{aligned} S_t &= \{\text{The scanner identifies a person as a terrorist}\} \\ S_u &= \{\text{The scanner identifies a person as an upstanding citizen}\} \\ P_t &= \{\text{The person being scanned is a terrorist}\} \\ P_u &= \{\text{The person being scanned is an upstanding citizen}\} \\ T_1 &= \{\text{The first person who scans positive is the terrorist}\} \end{aligned}$$

From the given data, we have the following probabilities:

$$\mathbb{P}[S_t|P_t] = 0.95 \quad \mathbb{P}[S_u|P_u] = 0.95$$

The probability that a person who scans positive is actually a terrorist is given by the posterior $\mathbb{P}[P_t|S_t]$; for this purpose, we use Bayes' rule in conjunction with the law of total probability:

$$\begin{aligned} \mathbb{P}[P_t|S_t] &= \frac{\mathbb{P}[S_t|P_t]\,\mathbb{P}[P_t]}{\mathbb{P}[S_t|P_t]\,\mathbb{P}[P_t] + \mathbb{P}[S_t|P_u]\,\mathbb{P}[P_u]} \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \\ \boxed{\mathbb{P}[P_t|S_t] \simeq 0.16} \end{aligned}$$

An alternative approach

This approach interprets the question as "what is the probability that the *first* positive scan is a true positive" as opposed to "what is the probability that a positive scan is a true positive".
Suppose that the terrorist is the n^{th} person that is scanned ($1 \leq n \leq 100$) and define this as the event S_n. For the 1^{st} person who scans positive to be the terrorist in this case, we need the scanner to produce $n-1$ true negatives followed by a true positive. Assuming that successive scans are independent, we can write:

$$\begin{aligned} \mathbb{P}[T_1|S_n] &= \mathbb{P}[S_u|P_u]^{n-1} \cdot \mathbb{P}[S_t|P_t] \\ &= (0.95)^n \end{aligned}$$

$$\mathbb{P}[T_1] = \sum_{n=1}^{100} \mathbb{P}[T_1 | S_n] \mathbb{P}[S_n]$$

Finally, we assume that all orderings of passengers are equally likely, leading to $\mathbb{P}[S_n] = 0.01$ for $1 \leq n \leq 100$ and we can write:

$$\begin{aligned}
\mathbb{P}[T_1] &= 0.01 \sum_{n=1}^{100} (0.95)^n \\
&= 0.01 \cdot \left[\frac{0.95\,(1 - 0.95^{100})}{1 - 0.95} \right] \\
\boxed{\mathbb{P}[T_1] \simeq 0.19}
\end{aligned}$$

To find the probability of the scanner scanning a person as a terrorist, we use the law of total probability:

$$\mathbb{P}[S_t] = \mathbb{P}[S_t | P_u] \mathbb{P}[P_u] + \mathbb{P}[S_t | P_t] \mathbb{P}[P_t]$$

Now, since only one person out of 100 is a terrorist, we have $\mathbb{P}[P_u] = 0.99$ and $\mathbb{P}[P_t] = 0.01$ so we get:

$$\begin{aligned}
\mathbb{P}[S_t] &= 0.05 \times 0.99 + 0.95 \times 0.01 \\
&= 0.059
\end{aligned}$$

Since this can be considered as the "success rate" of 100 Bernoulli trials, the expected number of successes is 100 times the success rate.

$$\begin{aligned}
\mathbb{E}[S_t] &= 100 \times 0.059 \\
\boxed{\mathbb{E}[S_t] \simeq 5.9}
\end{aligned}$$

Since the number of positive scans must be an integer we should expect about 6 positive scans.

Problem 2

Since there are no constraints placed on $\mathbb{P}[x]$, the maximum likelihood estimate of λ must be such that the expected value of x is set to the sample mean (BRML, chapter 8). We proceed as follows:

$$\begin{aligned}
\mathbb{E}[x] &= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda} \lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{e^{-\lambda} \lambda^{x-1} \cdot \lambda}{x\,(x-1)!} \qquad \text{(first term vanishes)} \\
&= \lambda \cdot \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!}
\end{aligned}$$

$$= \lambda \cdot \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{(y)!} \qquad (y = x-1)$$
$$= \lambda$$

where the last step follows from the normalization of the Poisson distribution[1]. Thus, the maximum likelihood estimator of λ $\left(\widehat{\lambda}_{\mathrm{ML}}\right)$ is given by setting it to the sample mean:

$$\boxed{\widehat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Now we set about proving that this actually maximizes the likelihood. Define $\chi = \{x_1, x_2, \cdots, x_n\}$ as the set of samples and $S = \sum_{i=1}^{n} x_i$. The likelihood is thus given by:

$$\begin{aligned} L &= \mathbb{P}\left[\chi \,|\lambda\right] \\ &= \prod_{i=1}^{n} \mathbb{P}\left[x_i \,|\lambda\right] \qquad \text{(by the i.i.d sampling assumption)} \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{e^{-n\lambda}\lambda^{S}}{\prod_{i=1}^{n} x_i!} \end{aligned}$$

We thus have:

$$\frac{dL}{d\lambda} = \frac{1}{\prod_{i=1}^{n} x_i!}\left[-ne^{-n\lambda}\lambda^{S} + e^{-n\lambda}S\lambda^{S-1}\right]$$

The λ estimate that maximizes likelihood causes the above derivative to vanish, so we have:

$$e^{-n\widehat{\lambda}_{\mathrm{ML}}}\left[S\left(\widehat{\lambda}_{\mathrm{ML}}\right)^{S-1} - n\left(\widehat{\lambda}_{\mathrm{ML}}\right)^{S}\right] = 0$$
$$\left(\widehat{\lambda}_{\mathrm{ML}}\right)^{S-1} e^{-n\widehat{\lambda}_{\mathrm{ML}}}\left[S - n\widehat{\lambda}_{\mathrm{ML}}\right] = 0$$

Since $\widehat{\lambda}_{\mathrm{ML}} = 0$ is a trivial solution (corresponds to $\mathbb{P}\left[x\right] \equiv 0 \,\forall x$), we have

$$\widehat{\lambda}_{\mathrm{ML}} = \frac{S}{n} \qquad \text{(Q. E. D.)}$$

Problem 3

1. Considering each classification as a Bernoulli trial with success rate p, observing a particular sequence of classifications with R correct ones and W incorrect ones corresponds to carrying out $R + W$ independent Bernoulli trials and observing R

[1] This normalization can be seen from the Maclaurin series expansion of the exponential function: $e^{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^{y}}{y!}$

successes. Moreover, when the data is generated purely at random, we have $p = \mathbb{P}[\text{correct classification}] = \mathbb{P}[\text{incorrect classification}] = 0.5$. We thus have

$$\begin{aligned}\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}] &= p^R (1-p)^W \\ &= 0.5^{R+W} \qquad \text{(Q. E. D.)}\end{aligned}$$

2. We have $\mathbb{P}[\mathcal{D}\,|\theta] = \theta^W (1-\theta)^R$ (note that there is a typo in the textbook). We can thus write

$$\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{non random}}] = \int_\theta \mathbb{P}[\mathcal{D}\,|\theta]\,\mathbb{P}[\theta]\,d\theta$$

For a Beta prior, we have:

$$\begin{aligned}\mathbb{P}[\theta] &= B(\theta\,|a,b) \\ &= \frac{1}{B(a,b)}\,\theta^{a-1}(1-\theta)^{b-1} \qquad 0 \le \theta \le 1\end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function. We can thus write

$$\begin{aligned}\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}] &= \frac{1}{B(a,b)}\int_0^1 \theta^W (1-\theta)^R \cdot \theta^{a-1}(1-\theta)^{b-1}\,d\theta \\ &= \frac{1}{B(a,b)}\int_0^1 \theta^{W+a-1}(1-\theta)^{R+b-1}\,d\theta \\ &= \frac{B(W+a,R+b)}{B(a,b)} \qquad \text{(Q. E. D.)}\end{aligned}$$

Note that, due to the aforementioned typo in the book replacing R with W causes the roles of a and b to be interchanged in this result as well. This won't make a difference to the final results since we assume a uniform prior in which $a = b = 1$.

3. For this part we make use of Bayes' rule and the law of total probability to write

$$\mathbb{P}[\mathcal{H}_{\text{random}}\,|\mathcal{D}] = \frac{\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}]\,\mathbb{P}[\mathcal{H}_{\text{random}}]}{\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}]\,\mathbb{P}[\mathcal{H}_{\text{random}}] + \mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{non random}}]\,\mathbb{P}[\mathcal{H}_{\text{non random}}]}$$

Since we consider the two hypotheses to be equally likely *a priori*, we have $\mathbb{P}[\mathcal{H}_{\text{random}}] = \mathbb{P}[\mathcal{H}_{\text{random}}]$ and

$$\begin{aligned}\mathbb{P}[\mathcal{H}_{\text{random}}\,|\mathcal{D}] &= \frac{\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}]}{\mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{random}}] + \mathbb{P}[\mathcal{D}\,|\mathcal{H}_{\text{non random}}]} \\ &= \frac{0.5^{R+W}}{0.5^{R+W} + \frac{B(W+a,R+b)}{B(a,b)}} \qquad \text{(Q. E. D.)}\end{aligned}$$

4. For the numerical evaluations in this part, we make use of the fact that for positive integer arguments, the gamma function can be expressed in terms of a factorial and

write:

$$\begin{aligned} B(a,b) &\equiv \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \\ &= \frac{(a-1)!\,(b-1)!}{(a+b-1)!} \qquad \text{(for } a,b \text{ positive integers)} \end{aligned}$$

Using this, we have

$$\begin{aligned} \mathbb{P}\left[\mathcal{H}_{\text{random}}\,|R=10,W=12\right] &= \frac{0.5^{22}}{0.5^{22}+\frac{B(13,11)}{B(1,1)}} \\ &= \frac{0.5^{22}}{0.5^{22}+\frac{12!10!}{23!}} \qquad \left(B(1,1)=\frac{0!\cdot 0!}{1!}=1\right) \end{aligned}$$

we thus have

$$\boxed{\mathbb{P}\left[\mathcal{H}_{\text{random}}\,|R=10,W=12\right] \simeq 0.78}$$

We can use this to compute Bayes' factor:

$$\begin{aligned} \frac{\mathbb{P}\left[\mathcal{H}_{\text{random}}\,|\mathcal{D}\right]}{\mathbb{P}\left[\mathcal{H}_{\text{non random}}\,|\mathcal{D}\right]} &= \frac{\mathbb{P}\left[\mathcal{H}_{\text{random}}\,|\mathcal{D}\right]}{1-\mathbb{P}\left[\mathcal{H}_{\text{random}}\,|\mathcal{D}\right]} \\ &\simeq \frac{0.78}{1-0.78} \\ &\simeq 3.55 \end{aligned}$$

This constitutes evidence for the fact that the data is randomly generated. Similarly,

$$\begin{aligned} \mathbb{P}\left[\mathcal{H}_{\text{random}}\,|R=100,W=120\right] &= \frac{0.5^{220}}{0.5^{220}+\frac{B(121,101)}{B(1,1)}} \\ &= \frac{0.5^{220}}{0.5^{220}+\frac{120!100!}{221!}} \\ &= \frac{0.5^{220}}{0.5^{220}+\frac{100!}{\prod_{i=121}^{221} i}} \end{aligned}$$

We thus have

$$\boxed{\mathbb{P}\left[\mathcal{H}_{\text{random}}\,|R=100,W=120\right] \simeq 0.83}$$

and Bayes' factor is similarly computed to be

$$\begin{aligned} \mathbb{P}\left[\mathcal{H}_{\text{random}}\,|R=100,W=120\right] &= \frac{0.5^{220}}{0.5^{220}+\frac{B(121,101)}{B(1,1)}} \\ &\simeq \frac{0.83}{1-0.83} \\ &\simeq 4.88 \end{aligned}$$

which constitutes even stronger evidence in favor of the hypothesis that the data was randomly generated. The conclusion, therefore, is that this classifier isn't better than random guessing.

5. A random classifier is equally likely to classify correctly or incorrectly, thus for a total of $n = R + W$ test examples, the probability of k incorrect classifications is given by a binomial distribution with a success rate of 0.5:

$$\begin{aligned}
\mathbb{P}[W = k] &= \binom{n}{k} 0.5^n \\
\mathbb{E}[W] &= \sum_{k=0}^{n} \binom{n}{k} k \cdot 0.5^n \\
&= 0.5^n \sum_{k=1}^{n} \frac{n!}{(k-1)!\,(n-k)!} \\
&= 0.5^n n \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} \\
&= 0.5^n n \sum_{k=1}^{n} \binom{n-1}{k-1} \\
&= 0.5n \sum_{l=0}^{n-1} \binom{n-1}{l} \cdot 0.5^{n-1} \\
&= 0.5n \qquad \text{(from the normalization of the binomial distribution)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[W^2\right] &= \sum_{k=0}^{n} \binom{n}{k} k^2 0.5^n \\
&= 0.5n \sum_{k=1}^{n} \frac{k \cdot (n-1)!}{(k-1)!\,(n-k)!} \cdot 0.5^{n-1} \\
&= 0.5n \sum_{l=0}^{n-1} \binom{n-1}{l} \cdot (l+1) \cdot 0.5^{n-1} \\
&= 0.5n\,[1 + 0.5\,(n-1)] \qquad \text{(combining normalization with the result for } \mathbb{E}[W]) \\
&= 0.25n\,(n+1) \\
\sigma_W^2 &= \mathbb{E}\left[W^2\right] - (\mathbb{E}[W])^2 \\
&= 0.25n\,(n+1) - 0.25n^2 \\
&= 0.25\,(R+W) \\
\sigma_W &= 0.5\sqrt{R+W} \qquad \text{(Q. E. D.)}
\end{aligned}$$

This means that, for a larger dataset, the variance of the number of errors is larger which means that we can observe more errors and still be confident in the hypothesis that the data is randomly generated. This means that inference is easier and it can be

seen from the above computation where the larger dataset gives more confidence in the hypothesis $\mathcal{H}_{\text{random}}$.

Problem 4

Let's define $f\left(c^{\text{pred}}\right)$ as the expected utility conditional on the predicted class:

$$\begin{aligned}
f\left(c^{\text{pred}}\right) &= \mathbb{E}\left[U \left| c^{\text{pred}} \right.\right] \\
&= \sum_{i=1}^{3} U\left(c^{\text{true}}, c^{\text{pred}}\right) \mathbb{P}\left[c^{\text{true}} = i \left| c^{\text{pred}} \right.\right] \\
&= \sum_{i=1}^{3} U\left(c^{\text{true}}, c^{\text{pred}}\right) \mathbb{P}\left[c^{\text{true}} = i \left| x \right.\right] \qquad \left(\text{knowing } x \text{ is equivalent to knowing } c^{\text{pred}}\right) \\
&= \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 5 & 3 & 1 \\ 0 & 4 & -2 \\ -3 & 0 & 10 \end{bmatrix} \\
&= \begin{bmatrix} 3.2 & 2.9 & 1.3 \end{bmatrix}
\end{aligned}$$

It is thus seen that the highest expected utility occurs for $c^{\text{pred}} = 1$. Thus, the best decision to make is $c^{\text{pred}} = 1$

Problem 5

Using Bayes' rule and the law of total probability, we can write

$$\begin{aligned}
\mathbb{P}\left[c = 1 \left| x \right.\right] &= \frac{\mathbb{P}\left[x \left| c = 1 \right.\right] \mathbb{P}\left[c = 1\right]}{\mathbb{P}\left[x \left| c = 1 \right.\right] \mathbb{P}\left[c = 1\right] + \mathbb{P}\left[x \left| c = 2 \right.\right] \mathbb{P}\left[c = 2\right]} \\
&= \frac{\mathbb{P}\left[x \left| c = 1 \right.\right]}{\mathbb{P}\left[x \left| c = 1 \right.\right] + \mathbb{P}\left[x \left| c = 2 \right.\right]} \qquad \left(\because \mathbb{P}\left[c = 1\right] = \mathbb{P}\left[c = 2\right] = 0.5\right) \\
&= \frac{1}{1 + \frac{\mathbb{P}[x|c=2]}{\mathbb{P}[x|c=1]}} \\
&= \frac{1}{1 + \frac{1/\sqrt{2\pi\sigma^2}\exp\left[-(x-m_2)^2/(2\sigma^2)\right]}{1/\sqrt{2\pi\sigma^2}\exp\left[-(x-m_1)^2/(2\sigma^2)\right]}} \\
&= \frac{1}{1 + e^{-\left(x^2 - 2m_2x + m_2^2 - x^2 + 2m_1x - m_1^2\right)/(2\sigma^2)}} \\
&= \frac{1}{1 + e^{-\left(\frac{m_1 - m_2}{\sigma^2}x + \frac{m_2^2 - m_1^2}{2\sigma^2}\right)}}
\end{aligned}$$

which fits the required form with

$$\boxed{a = \frac{m_1 - m_2}{\sigma^2}} \qquad \boxed{b = \frac{m_2^2 - m_1^2}{2\sigma^2}}$$