1. Bayes' Rule for medical diagnosis

Solution:

\[
\begin{align*}
p(r = 1 | s = 1) &= 0.99 & p(r = 0 | s = 1) &= 0.01 \\
p(r = 1 | s = 0) &= 0.01 & p(r = 1 | s = 1) &= 0.99
\end{align*}
\]

where \( r = 1 \) indicates the test is positive, and \( s = 1 \) indicates that you are sick.

\[
p(s = 1 | r = 1) = \frac{p(s = 1, r = 1)}{p(r = 1)} = \frac{0.99 \cdot \frac{1}{10000}}{0.99 \cdot \frac{1}{10000} + 0.01 \cdot \frac{9999}{10000}} = \frac{1}{102}
\]

2. The Monty Hall Problem

Solution:

Let \( H_i \) denote the hypothesis that the prize is behind door \( i \). We make the following assumptions: the three hypotheses \( H_1, H_2, H_3 \) are equally likely a priori.

\[ P(H_1) = P(H_2) = P(H_3) = \frac{1}{3} \]

WLOG we choose door 1. The data we receive after choosing door 1 is either door 3 or door 2 is opened (e.g. \( D = 3 \) or \( D = 2 \)). We can compute the conditional likelihood of the data given our hypothesis is:

\[
\begin{align*}
p(D = 1 | H_1) &= 0 & p(D = 1 | H_2) &= 0 & p(D = 1 | H_3) &= 0 \\
p(D = 2 | H_1) &= \frac{1}{2} & p(D = 2 | H_2) &= 0 & p(D = 2 | H_3) &= 1 \\
p(D = 3 | H_1) &= \frac{1}{2} & p(D = 3 | H_2) &= 1 & p(D = 3 | H_3) &= 0
\end{align*}
\]

And using Bayes rule:

\[
P(H_i | D = 3) = \frac{P(D = 3 | H_i)P(H_i)}{P(D = 3)}
\]

We can see that \( P(D = 3) \) is \( 1/2 \) by renormalizing the conditional likelihood and adding up the individual instances of \( P(D = 3 | H_i) \) so

\[
p(H_1 | D = 3) = \frac{1}{3} \quad p(H_2 | D = 3) = \frac{2}{3} \quad p(H_3 | D = 3) = 0
\]

Thus you should always switch to door 2 after choosing door 1 and seeing door 3 opened.

3. Conditional Independence
Solution:

a) Bayes’ rule gives

\[ P(H|e_1, e_2) = \frac{P(e_1, e_2|H)p(H)}{P(e_1, e_2)} \]

We can conclude that the information in (ii) is sufficient for the calculation.

For (i) we know \( P(e_1, e_2) \) and \( P(H) \), but we do not know \( P(e_1, e_2|H) \). However, we do not know if \( P(e_1) \perp P(e_2)|H \) so we cannot calculate \( P(e_1, e_2|H) \), and thus we cannot calculate \( P(H|e_1, e_2) \).

Likewise, we cannot calculate \( P(e_1, e_2) \) for (iii), so (iii) is not sufficient.

b) Now the equation simplifies to

\[ P(H|e_1, e_2) = \frac{P(e_1|H)P(e_2|H)p(H)}{P(e_1, e_2)} \]

so (i) and (ii) are obviously sufficient. (iii) is also sufficient because we can compute \( P(e_1, e_2) \) using normalization.

4. Irrelevant features with Naive Bayes

Solution:

a) By Bayes rule, we have

\[ \frac{p(c = 1|x)}{p(c = 2|x)} = \frac{p(x|c = 1)p(c = 1)}{p(x|c = 2)p(c = 2)} = \frac{p(x|c = 1)}{p(x|c = 2)} \]

which is just the likelihood ratio. Hence

\[ \log \frac{p(c = 1|x)}{p(c = 2|x)} = \phi(x_i)^T(\beta_1 - \beta_2) \]

b) If \( \beta_{1,w} = \beta_{2,w} \) then \( \beta_{1,w} - \beta_{2,w} = 0 \), such that word \( w \) will be ignored. This is equivalent to requiring the log odds ratio, \( \log \frac{\hat{\theta}_{1,w}}{1-\hat{\theta}_{1,w}} \), to be the same for both classes. If \( \theta_{1,w} = \theta_{2,w} \) we can safely remove the word without changing the log posterior odds ratio.

c) The estimates would be

\[ \hat{\theta}_{1,w} = \frac{1 + n_1}{2 + n_1} \quad \hat{\theta}_{2,w} = \frac{1 + n_2}{2 + n_2} \]

We see that these are different, since \( n_1 \neq n_2 \). Hence \( \frac{\theta_{1,w}}{1-\theta_{1,w}} \neq \frac{\theta_{2,w}}{1-\theta_{2,w}} \) so the word will not be ignored. If we use the MLE, we avoid this problem, however the MLE results in overfitting.

d) We can use feature selection methods to try to remove irrelevant words. This is similar to imposing the prior that says \( \theta_{c,w} \) is the same across classes.

5. Fitting a naive Bayes spam filter by hand
6. Which of the following problems are more suited for a learning approach and which are more suited for a design approach?

**Solution:**

1) Either. Design approach will allow for experimental design and higher certainty but if data is available it may be faster and quicker to use a Learning approach.

2) Design: Math has ways of checking if a number is prime though finding prime numbers is slightly more challenging. Additionally, number theory suggests there might not be a pattern in terms of when prime numbers occur.

3) Learning: Data exists, there’s no clear analytical solution

4) Design: Physics has theories and laws that can easily and accurately predict this.

5) Learning: data is cheap, system is not easily represented as an equation that has an optimal solution, there is likely a pattern in the data.

7. You have an unfair 6-sided die (i.e. 1 dice). Values 1 and 4 are rolled with a probability 1/4. All other values are rolled with probability 1/8

**Solution:**

a) 

\[ E[x] = \sum_{i=1}^{m} x_i p(x_i) \]

where \( x_i \) can take discrete values 1 to m.

b) The expected value of the weighted dice is

\[
\sum_{d=1}^{6} Pr(d) \ast d = 1 \frac{1}{4} + 2 \frac{1}{8} + 3 \frac{1}{8} + 4 \frac{1}{4} + 5 \frac{1}{8} + 6 \frac{1}{8} = \frac{26}{8} = 3.25
\]

c) 

\[
\sum_{d=1}^{6} Pr(d)(d - 3.25)^2 = \\
\frac{1}{4}((1 - 3.25)^2 + (4 - 3.25)^2) + \frac{1}{8}((2 - 3.25)^2 + (3 - 3.25)^2 + (5 - 3.25)^2 + (6 - 3.25)^2) = \frac{47}{16}
\]
Solution: There are two approaches Bayesian estimation and MLE.

The first solution is the MLE which assumes both $\mu$ and $\sigma$ are unknown. Here we can write down the normal density function, convert it to log-likelihood, compute derivative and set it equal to 0. Solving that equation will tell us the MLE for $\mu$ and $\sigma$ both.

We can also use Bayesian estimation where we assume $\mu$ is unknown but $\sigma$ is known.

Solution 1:

$$L(\mu, \sigma) = f(x|\mu, \sigma) = \prod_{i=1}^{n}(2\pi\sigma)^2 \exp[-(x - \mu)^2 / 2\sigma^2]$$

$$\ell(\mu, \sigma) = \sum_{i=1}^{n}[- \log \sigma - 1/2 \log 2\pi - 1/2\sigma^2 (x_i - \mu)^2]$$

$$= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\sigma^2 \pi}} \exp\left[-(x_i - \mu)^2 / 2\sigma^2\right]\right)$$

$$= \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \log \sqrt{2\sigma^2 \pi}\right)$$

$$\frac{d \log L(x|\mu, \sigma)}{d\mu} = \sum_{i=1}^{n} -\frac{2 \cdot -1(x_i - \mu)}{2\sigma^2} = 0$$

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

$$\sum_{i=1}^{n} x_i = n \cdot \hat{\mu}$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \hat{\mu}$$

Solution 2

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

also the Gaussian prior has a mean distribution that is itself normal: $p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(X)} = \alpha \times \prod_{i}^{n} p(x_i|\mu)p(\mu)$$

and $\alpha = \frac{1}{p(X)}$
Note: $\alpha$ is independent of $\mu$.

$$p(\mu|x) = \alpha \times \prod_{i}^{n} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x_i - \mu)^2}{2\sigma^2}]$$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} \frac{-(\mu - \mu_0)^2}{2\sigma_0^2}$$

$$\rightarrow \beta \times \exp \left[ -\frac{1}{2} \left( \sum_{i}^{n} \left( \frac{-(x_i - \mu)^2}{\sigma^2} + \frac{-(\mu - \mu_0)^2}{\sigma_0^2} \right) \right) \right]$$

$$\rightarrow \gamma \times \exp \left[ -\frac{1}{2}[n(\sigma^2 + \sigma_0^2)]^2 - 2(1/\sigma^2 \sum_{i}^{n} x_i - \mu_0/\sigma_0^2)\mu + \gamma = \frac{(\mu - \mu_n)^2}{\sigma_n} \right]$$

So from $\mu_n$ and $\sigma_n^2$:

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \rightarrow \mu_n = \sum_{i}^{n} x_i / n = \bar{X}$$

So our MLE for $\mu$ would (again) be $\frac{1}{n} \sum_{i=1}^{n} x_i$

9. Bias-variance decomposition (from ISLR)

Solution:

b) Squared bias is large in simple models, as flexibility increases the bias decreases as the model is more able to approximate the true underlying function. If bias is too high, the model is likely to underfit.

Variance increases as the model flexibility increases as the model is more able to account for weaker trends in the training data. One way to think about variance is a quantification of how the model would change if it was constructed using a different training data sample. If
the model would change significantly, then the model is a high variance model and has likely overfit the training data.

Training error monotonically decreases as flexibility increases as more and more specific trends are being captured by the model. Testing data takes a u-shape where the left side is under-fitting, and the right side is over-fitting. Testing error is the sum of bias-squared plus variance.

10. From FML chapter 2. Let $\mathcal{X} = \mathbb{R}^2$. Consider the set of concepts of the form $c = (x, y) : x^2 + y^2 \leq r^2$ for some real number $r$. Show that this class can be $(\epsilon, \delta)$-PAC-learned from the training data of size $m \geq (1/\epsilon)\log(1/\delta)$.

**Solution:** WLOG, suppose our target concept $c$ is the circle around the origin with radius $r$. We will choose a slightly smaller radius 4 by

$$s := \inf\{s' : P(s' \leq ||x|| \leq r) < \epsilon\}$$

Let $A$ denote the annulus between radii $s$ and $r$; that is, $A := \{x : s \leq ||x|| \leq r\}$

By definition of $s$,

$$P(A) \geq \epsilon$$

In addition, our generalization error, $P(c\Delta L(T))$, must be small if $T$ intersects $A$. We can state this as

$$P(c\Delta L(T)) > \epsilon \implies T \cap A = \emptyset$$

Using $P(A) \geq \epsilon$, we know that any point in $T$ chosen according to $p$ will “miss” region $A$ with probability at most $1 - \epsilon$. Defining error := $P(c\Delta L(T))$, we can combine this with the above equation to see that

$$P(\text{error} > \epsilon) \leq P(T \cap A = \emptyset) \leq (1 - \epsilon)^m \leq e^{-m\epsilon}$$

Setting $\delta$ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$