CSE 515: Statistical Methods in Computer Science

Homework Assignment 1

Chenglong Wang

1 Probability

Solution for Problem 1.

The chance of actually having the disease given positive test result is $P(disease|positive)$.

$$\begin{aligned}
P(disease \mid positive) &= \frac{P(positive \mid disease) \times P(disease)}{P(positive)} \\
&= \frac{P(positive \mid disease) \times P(disease)}{P(positive \mid disease) \times P(disease) + P(positive \mid \neg disease) \times P(\neg disease)} \\
&= \frac{0.98 \times 10^{-4}}{0.98 \times 10^{-4} + 0.02 \times (1 - 10^{-4})} \\
&= 0.0049
\end{aligned}$$

Solution for Problem 2.

1. Suppose you reach into the bag, pick out a coin at random, flip it and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

$$\begin{aligned}
P(fake \mid head) &= \frac{P(head \mid fake) \times P(fake)}{P(head)} \\
&= \frac{P(head \mid fake) \times P(fake)}{P(head \mid fake) \times P(fake) + P(head \mid \neg fake) \times P(\neg fake)} \\
&= \frac{1 \times \frac{1}{n}}{1 \times \frac{1}{n} + 0.5 \times \frac{n-1}{n}} \\
&= \frac{2}{n+1}
\end{aligned}$$

2. Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

$$\begin{aligned}
P(fake \mid k_heads) &= \frac{P(k_heads \mid fake) \times P(fake)}{P(k_heads)} \\
&= \frac{P(k_heads \mid fake) \times P(fake)}{P(k_heads \mid fake) \times P(fake) + P(k_heads \mid \neg fake) \times P(\neg fake)} \\
&= \frac{1 \times \frac{1}{n}}{1 \times \frac{1}{n} + 0.5^k \times \frac{n-1}{n}} \\
&= \frac{2^k}{n + 2^k - 1}
\end{aligned}$$

3. Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure retures fake if all k flips come up heads; otherwise it returns normal. What is the (unconditional) probability that this procedure makes an error?

$$\begin{aligned} P(error) &= P(fake, \neg k\_heads) + P(\neg fake, k\_heads) \\ &= 0 + P(k\_heads \mid \neg fake) \times P(\neg fake) \\ &= \frac{n-1}{2^k \cdot n} \end{aligned}$$

2 Conditional Independence

Solution for Problem 3.

(a) **Weak Union:**

1. Given $(X \perp Y, W|Z)$, we have $P(X, Y, W|Z) = P(X|Z)P(Y, W|Z)$, by summing over all values of Y, we have $P(X, W|Z) = P(X|Z)P(W|Z)$. According to Beysian rule, we have $P(X|W, Z)P(W|Z) = P(X|Z)P(W|Z)$, which equals to have the fact that $P(X|W, Z) = P(X|Z)$.
2. Using Beysian rule, we have $P(X, Y|Z, W) = P(X|Y, Z, W)P(Y|Z, W)$.
3. Using the property $(X \perp Y, W|Z)$, we have:

$$\begin{aligned} (X \perp Y, W|Z) \Longrightarrow & P(X, Y, W|Z) = P(X|Z)P(Y, W|Z) \\ \Longrightarrow & P(X|Y, W, Z) \cdot P(Y, W|Z) = P(X|Z) \cdot P(Y, W|Z) \end{aligned}$$

4. (Goal) We need to prove $(X \perp Y|Z, W)$, which is same as $P(X, Y|Z, W) = P(X|Z, W)P(Y|Z, W)$. To achieve this goal, using fact 2, we only need to prove $P(X|Y, Z, W) = P(X|Z, W)$. To prove this goal, using fact 3, we only need to prove $P(X|Z) = P(X|Z, W)$ if $P(Y, W|Z) \neq 0$ (if $P(Y, W|Z) = 0$, the original goal is trivial since both sides are 0). Since fact 1 proves this, our proof goal is achieved.

(b) **Contraction:**

We have the following two facts.

1. Given $(X \perp W|Z, Y)$, we have $P(X, W|Z, Y) = P(X|Z, Y)P(W|Z, Y)$.
2. Given $(X \perp Y|Z)$, we have $P(X, Y|Z) = P(X|Z)P(Y|Z)$.

Our goal can be transformed in the following way:

- In order to prove $(X \perp Y, W|Z)$, we need to prove $P(X, Y, W|Z) = P(X|Z)P(Y, W|Z)$.
- Since $P(X, Y, W|Z) = P(X, W|Y, Z)P(Y|Z)$, we only need to prove that $P(X, W|Y, Z)P(Y|Z) = P(X|Z)P(Y, W|Z)$.
- According to fact 1, we only need to prove $P(X|Z, Y)P(W|Z, Y)P(Y|Z) = P(X|Z)P(Y, W|Z)$.
- Sicne $P(Y, W|Z) = P(Y|W, Z)P(W|Z)$, we only need to prove $P(X|Z, Y)P(W|Z, Y)P(Y|Z) = P(X|Z)P(Y|W, Z)P(W|Z)$. This goal can be simplified to prove $P(X|Z, Y) = P(X|Z)$.

Given fact 2 that $P(X, Y|Z) = P(X|Z)P(Y|Z)$, we have $P(X|Y, Z)P(Y|Z) = P(X|Z)P(Y|Z)$, we can derive our goal if $P(Y|Z) \neq 0$. Otherwise if $P(Y|Z) = 0$, the original goal is immediately satisfied. These two facts finishes the proof.

(c) **Intersection:**

We have the following facts.

1. $(X \perp Y|Z, W)$ indicates that $P(X, Y|Z, W) = P(X|Z, W)P(Y|Z, W)$.
2. $(X \perp W|Z, Y)$ indicates that $P(X, W|Z, Y) = P(X|Z, W)P(W|Z, Y)$.
3. Combining the two facts above, we have:

$$\frac{P(X, Y|Z, W)}{P(Y|Z, W)} = \frac{P(X, W|Z, Y)}{P(W|Z, Y)}$$

4. The fact above equals to $P(X, Y|Z, W)P(W|Z, Y) = P(Y|Z, W)P(X, W|Z, Y)$.
5. Summing over the W on the formula above, we have $P(X, Y|Z) = P(Y|Z)P(X|Z, Y)$ and it is equivalent to have $P(X|Z) = P(X|Z, Y)$. Since $P(X|Z, Y) = \frac{P(X,Y|Z)}{P(Y|Z)}$, we have $P(X, Y|Z) = P(Y, Z)P(X, Z)$, which indicates $(X \perp Y|Z)$.

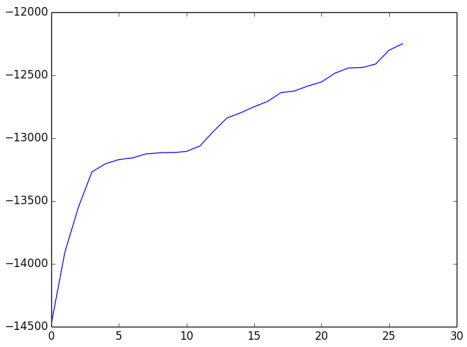
According to the contraction property and $(X \perp Y|Z)$ and $(X \perp W|Z, Y)$, we can prove the problem.

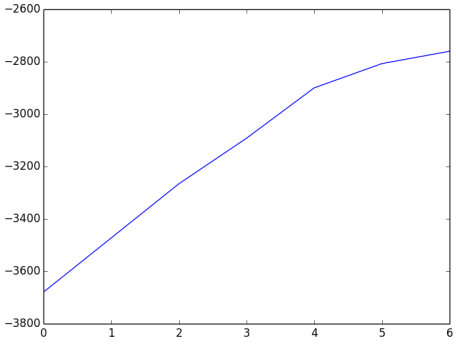
(d) **Counterexample:**

Consider a distribution where $X = Y = W$, and their values are independent of the value of Z. Then the two conditions of the intersection property holds: $(X \perp W|Z, Y)$ and $(X \perp Y|Z, W)$. However, it is not the case that $P(X, Y, W|Z) \neq P(X|Z)P(Y, W|Z)$ because Z cannot determine the cases for $X = Y = W = 1$ and $X = Y = W = 0$.

3 Programming: EM for Mixtures of Gaussians

(a). For the traning data (left), it takes 27 iterations to converge and for the test data (right) it takes only 6 iterations.





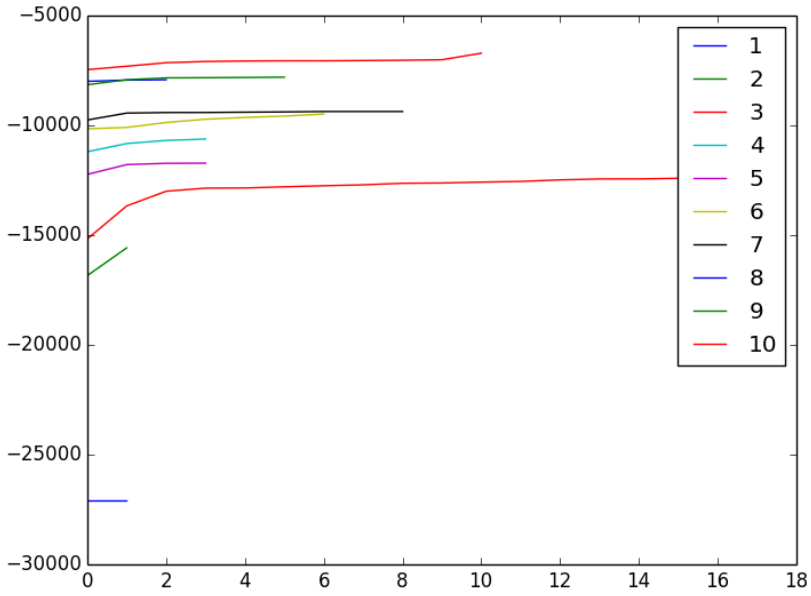
Note that I didn't include the initial likelihood in the grammar since the first iteration has a big jump, which makes the rest changes indistinguishable in the figure.

(b). In my solution for this first type of randomization, I pick μ_k for each feature k is sample from a uniform distribution between ranges $[min_k, max_k]$, where min_k and max_k are min-max values for the k-th feature of all data in the dataset. With this way of sampling, two methods works roughly the same. The reason is that both methods similar chances to hit the real means for each class.

(c). When I ran 10 times, the log likelihood stays pretty similar. The values are -12255.503663, -12255.503663, -12442.4946406, -12255.503663, -12688.6646817, -12688.6646745, -12159.5228952, -12186.985494, -12688.6646745, -12254.5239628, -12121.9834305.

(d). Our algorithm is able to classify 103 out of 142 correct. Major faults are result from the difficulties in distinguishing the first and the third types of wines.

(e). The figure for different clustering is shown below.



I noticed that 3 clusters contains best likelihood, while given 1 is the worst. The reason for that is that having more or less clusters tend to make the model overfit to training data, and the result is less satisfying likelihood.