Korea Advanced Institute of Science and Technology
School of Electrical Engineering

**EE531 Statistical Learning Theory**

**Midterm**

13:00-15:45, Tuesday, Oct. 22 2019

---

- Only the answer booklet will be graded. Scratch paper is available but should not be handed in and will not be considered in the grading.
- There are 5 problems of varying difficulty use your time wisely.
- This exam is not meant to test your number crunching ability. Think before you do any calculations.

Good Luck!

**Problem 1** (20 pts)
**(T/F problem)** Please state whether the following statements are True or False. Explain your reasoning in one sentence.

(a) (2pts) As training data size increases, the empirical error rises and saturates while the test error drops and then plateaus. (**T** / F)

**[solution]** If we have a small number of training data, the empirical error will be very small as the model will fit the data very well. As the number of training data increases, the model will have a hard time fitting the data and the empirical error will rise. The test error will be initially high and as the model is trained with more data the error will fall and ultimately the test error will be close to the empirical error. Since the model complexity is fixed, both error will saturate.

(b) (2pts) Misleading sample set is a sample set that is accurately predicted by a hypothesis with expected 0-1 loss greater than some $\epsilon$. The set of samples that make the expected 0-1 loss of the empirical risk minimizer greater than $\epsilon$ is a subset of the misleading sample set. (**T** / F)

**[solution]** Refer from the lecture 2, The bad hypothesis $\mathcal{H}_B$ is defined by

$$\mathcal{H}_B = \{f_\theta \in \mathcal{H} : \mathcal{L}_{(D,f)}(f_\theta) > \epsilon\}, \tag{1}$$

and misleading samples $\mathcal{M}$ is defined by

$$\mathcal{M} = \{S|_x : \exists f \in \mathcal{H}_B, \mathcal{L}_S(f) = 0\}. \tag{2}$$

Then, $S|_x : \mathcal{L}_{(D,f)}(f_s) > \epsilon \subseteq \mathcal{M}$.

(c) (2pts) A PAC learnable hypothesis space must be of finite size and must have a hypothesis with zero test error, and an agnostic PAC learnable hypothesis space is of infinite size and assumes an existence of a joint distribution over the data and label. (T / **F**)

**[solution]** There is no condition regarding the size of hypothesis space for deciding whether the hypothesis space is PAC learnable or agnostic PAC learnable.

(d) (2pt) Assume that a training set $\mathcal{S}$ is $\epsilon$-representative w.r.t. domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function $l$, and distribution $D$. Then any output of $\text{ERM}_\mathcal{H}(\mathcal{S})$, namely, any $f_\mathcal{S} \in \arg\min_{f\in\mathcal{H}} \mathcal{L}_\mathcal{S}(f)$, satifies

$$\mathcal{L}_D(f_\mathcal{S}) \leq \min_{f\in\mathcal{H}} \mathcal{L}_D(f) + \epsilon.$$

(T / **F**)

**[solution]** The training set $S$ should be $\frac{\epsilon}{2} - representative$ for satisfying the inequality.

(e) (2pts) No free lunch theorem states that without inductive bias PAC learning is impossible. (**T**/F)

**[solution]** The No Free Lunch theorem shows that PAC-learning is impossible without restricting the hypothesis class $\mathcal{H}$.

(f) (2pts) Test error is defined as the probability of prediction error for 0-1 loss. (**T** / F)

**[solution]** For the general case: $L_D(f_\theta) = \mathbb{E}_{(\mathbf{x},y)\sim D}[I(f_\theta(\mathbf{x}), y)] = P_{(\mathbf{x},y)\sim D}[f_\theta(\mathbf{x}) \neq y]$

(g) (2pts) Beta distribution which sometimes referred to as the probability of the probability is the conjugate prior of the Binomial distribution which is a distribution over the number of positive outcomes from a fixed number of trials. (**T**/F)

**[solution]** Refer to the lecture 5.

(h) (2pt) You are told that random variable $x$ has a range between 0 and $z$. The maximal likelihood estimate of $z$ is the largest observed value of $x$. (T/**F**)

**[solution]** It is true when $x$ is from uniform distribution, but it is not always true when we don't know from which distribution does $x$ come from. For example, let $x$ come from a triangular distribution whose peak value is at $z/2$, and suppose we observed two points $x_1$, $x_2$. Then, the maximum likelihood estimation of $z$ will be $x_1 + x_2$, which is greater than the largest observed value of $x$.

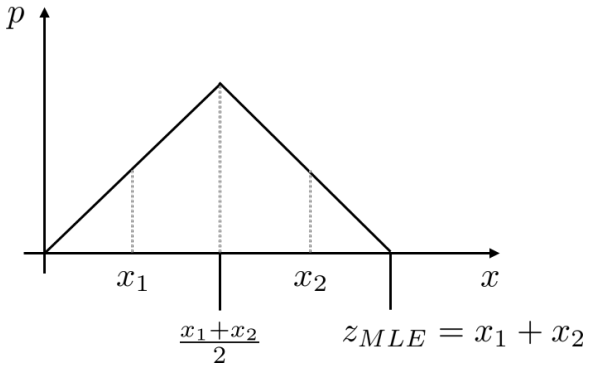


Figure 1: A counterexample where $x$ is drawn from a triangular distribution.

(i) (2pt) Expectation Maximization (EM) is a parameter estimation algorithm that iteratively evaluates the conditional probability of the latent variable given the observation and then maximizes the log likelihood over the conditional probability of the latent with respect to the parameter. (**T**/F)

**[solution]** Let denote C is the latent variable, D is observation, $\theta$ is a parameter. In the EM algorithm, at E-step, we compute $P(C|D, \theta)$. At M-step, we find $\hat{\theta}$ such that: $\hat{\theta} = \arg\max_{\bar{\theta}} \mathbb{E}_{P(C|D,\theta)}[P(C, D|\bar{\theta})]$. We repeat two steps until $\theta$ doesn't change much

(j) (2pts) For the three random variables, $A$, $B$ and $C$, when $A \perp\!\!\!\perp B|C$: (1) $P(A, B, C) = P(A)P(B|A)P(C|A, B)$, (2) $P(A, B) = P(A)P(B|A)$, (3) $P(A|B, C) = P(A|C)$. (**T** / F)

**[solution]** (1) and (2) is always true by the product rule. (3) is true since A and B is conditionally independent given C.

**Problem 2** (20 pts)

(a) **(PAC Learning)** Assume a particular plant will be "ok" if it is kept below the temperature $a$ such that temperature $\leq a$ where $a \in [0, 100]$. Assume a hypothesis set

$$\mathcal{H} = \{[0, a] | a \in [0, 100]\}.$$

Assume a simple algorithm: Observe $m$ temperatures the plant survived, and return $[0, b]$ where $b$ is the highest temperature the plant survived. How many examples do we need to have such that the error is within $\epsilon$ of the true value $a$ with confidence $(1 - \delta)$. Use the fact that $(1 - \epsilon) \leq e^{-\epsilon}$.

**[solution]**

Let $D$ is the distribution of the observed temperature $x$ that a plant is 'ok', $S$ is a training set containing $m$ observed temperature that the plant survived, formally, $S \sim D^m$. Let assume that $a^*$ is the true temperature such that the plant will be "ok", $a^* \in [0, 100]$, and $f$ be the corresponding hypothesis, i.e. $f = [0, a^*]$. Since we always observe the temperature that the plant is survived, hence, $x \in [0, a*]$.

The hypothesis space is $\mathcal{H} = \{[0, a] | a \in [0, 100]\}$. From the selected hypothesis space, the realizability assumption is hold.

Let's call the algorithm as stated in problem is $A(S) = [0, b]$, where $b = \max_i \{x_i\}_{i=1}^m, x_i \in S$. The definition of algorithm $A$ implies that $A(S) \subseteq [0, a^*]$ for every $S$. Thus, the error of prediction is

$$L_{(D,f)}(A(S)) = D(\{S|_x : x \in [0, a^*] \backslash [0, b]\})$$

Fix some $\epsilon \in (0, 1)$. Defining the temperature $a_0$ such that $P(x \in [0, a_0]) = 1 - \epsilon$, and $P(x \in [a_0, a^*]) = \epsilon$.

If $L_{(D,f)}(A(S)) > \epsilon$, then $b < a_0$. We obtain,

$$\begin{aligned} D^m(S : L_{(D,f)}(A(S)) > \epsilon) &= P(b < a_0) \\ &= P(\max_i \{x_i\}_{i=1}^m < a_0) \\ &= P(\{x_i\}_{i=1}^m < a_0) \\ &= \prod_{i=1}^m P(x_i < a_0) \\ &= \prod_{i=1}^m P(x_i \in [0, a_0]) \\ &= (1 - \epsilon)^m \leq e^{-m\epsilon} \end{aligned} \tag{3}$$

Since the error is within $\epsilon$ with confidence $(1 - \delta)$. Then we have,

$$D^m(S : L_{(D,f)}(A(S)) > \epsilon) < \delta \tag{4}$$

Combine (3) and (4), we have: $m \geq \frac{1}{\epsilon} \log(\frac{1}{\delta})$

(b) You are given a hypothesis space of size $k$, $\mathcal{H} = \{h_1(\cdot), \ldots, h_k(\cdot)\}$. Given $m$ training data $\mathcal{X} = \{(x_j, f(x_j))\}_{j=1}^m$, test loss $L_D(h_i(x), f(x)) = \mathbb{P}(h_i(x) \neq f(x))$ and training loss $L_s(h_i(x), f(x)) = \frac{1}{m} \sum_{j=1}^m \mathbb{I}(h_i(x_j) \neq f(x_j))$. Here $\mathbb{P}(\cdot)$ and $\mathbb{I}(\cdot)$ denote respectively probability and indicator function.

i Prove the following:

$$P(\forall h_i(x) \in \mathcal{H}, |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| < \epsilon) \geq 1 - 2ke^{-2\epsilon^2 m}.$$

Hint:

$$\mathbb{P}(\exists h_i(\cdot) \in \mathcal{H}, |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| > \epsilon) = \mathbb{P}(E_1 \cup \ldots \cup E_k)$$

where $E_i$ denotes event $|L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| > \epsilon$ and use the fact that $\mathbb{P}(E_i) \leq 2e^{-2\epsilon^2 m}$

[**Solution**]

$$\begin{aligned}
P(\forall h_i(x) \in \mathcal{H}, |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| < \epsilon) &= 1 - \mathbb{P}(\exists h_i(\cdot) \in \mathcal{H}, |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| \geq \epsilon) \\
&= 1 - P(\bigcup_{i=1}^{k} |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| \geq \epsilon) \\
&\geq 1 - \sum_{i=1}^{k} P(|L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| \geq \epsilon) \text{ (Union bound)} \\
&\geq 1 - \sum_{i=1}^{k} 2e^{-2\epsilon^2 m} \\
&= 1 - 2ke^{-2\epsilon^2 m}
\end{aligned}$$

ii Let $\mathcal{H}$ be the space of all pure conjunction formulas (in other words using only the operation AND where $x$ AND $y$ is true only when both $x$ and $y$ is true) over $n$ Boolean attributes. What is the least number of samples needed to guarantee the statement above in (i)?

[**Solution**]

Each attribute have three possible values including *true*, *false*, and *don't care*. Therefore, we have total $3^n$ hypotheses, in other words, $k = 3^n$. To guarantee the statement in (i), it means,

$$P(\forall h_i(x) \in \mathcal{H}, |L_D(h_i(x), f(x)) - L_s(h_i(x), f(x))| < \epsilon) \geq 1 - 2ke^{-2\epsilon^2 m} > 1 - \delta$$

Then we have,

$$2ke^{-2\epsilon^2 m} < \delta$$

Since $k = 3^n$, then the number of samples we need to guarantee is,

$$m \geq \frac{1}{2\epsilon^2} \ln(\frac{2.3^n}{\delta})$$

**Problem 3** (20 pts)
**(VC-Dimension)** Please verify the answer for following sub-questions.

(a) Consider the function $f$ in $\mathbb{R}^d$ given below. How many points can it shatter or more specifically what is its VC dimension? Explain your answer.

$$f(x) = \begin{cases} 1 & \text{if } W^T\mathbf{x} + b > 0, \\ -1 & \text{otherwise} \end{cases}$$

where $\mathbf{x} \in \mathbb{R}^d$, $W \in \mathbb{R}^{d\times 1}$, and $b \in \mathbb{R}$.

**[Solution]** It will be proved in two steps:

i **There exist $d+1$ points that the function $f$ can shatter.**
Consider $d+1$ points $\mathbf{x}^{(0)} = (0, ..., 0)^T$, $\mathbf{x}^{(1)} = (1, 0, ..., 0)^T$, $\mathbf{x}^{(2)} = (0, 1, ..., 0)^T$, $\cdots$, $\mathbf{x}^{(d)} = (0, 0, ..., 1)^T$. These $d+1$ points are arbitrarily labeled: $\mathbf{y} = (y_0, y_1, ..., y_d)^T \in \{-1, 1\}^{d+1}$.
Let $b = 0.5 \cdot y_0$ and $\mathbf{w} = (w_1, w_2, ..., w_d)$ where $w_i = y_i$, $i \in \{1, 2, ..., d\}$. Thus $f(\mathbf{x})$ can label all these $d+1$ points correctly.
So the VC dimension of the given function $f$ is at least $d+1$.

ii **No $d+2$ (or more) points can be shattered by the function $f$.**
Expand $\mathbf{x} \in \mathbb{R}^d$ to $\hat{\mathbf{x}} \in \mathbb{R}^{d+1}$, by letting $\hat{\mathbf{x}}^T = (\mathbf{x}^T, 1)$, and let $\hat{\mathbf{w}}^T = (\mathbf{w}^T, b)$. Thus:

$$f(\hat{\mathbf{x}}) = \begin{cases} 1 & \text{if } \hat{\mathbf{w}}^T\hat{\mathbf{x}} > 0, \\ -1 & \text{otherwise} \end{cases}$$

Assume there exist $d+2$ points that the function $f$ in $\mathbb{R}^d$ can shatter, namely $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(d+1)} \in \mathbb{R}^d$ corresponding to $\hat{\mathbf{x}}^{(0)}, \hat{\mathbf{x}}^{(1)}, \cdots, \hat{\mathbf{x}}^{(d+1)} \in \mathbb{R}^{d+1}$.
Since $d+2$ points are in $\mathbb{R}^{d+1}$, there exists certain $i$ s.t.:

$$\hat{\mathbf{x}}^{(i)} = \sum_{j\neq i} a_j \cdot \hat{\mathbf{x}}^{(j)},$$

where at least one $a_j \neq 0$. Let $S = \left\{j | j \neq i, a_j \neq 0\right\}$.
$\forall j \in S$, we give $\mathbf{x}^{(j)}$ the label $sign(a_j)$, and give $\mathbf{x}^{(i)}$ a label $-1$. By our assumption, there exists $\hat{\mathbf{w}}$ that make $f(\hat{\mathbf{x}})$ label those $d+2$ points correctly. So $\forall j \in S$, we have

$$a_j \cdot \hat{\mathbf{w}}^T\hat{\mathbf{x}}^{(j)} > 0,$$

and

$$\hat{\mathbf{w}}^T\hat{\mathbf{x}}^{(i)} \leq 0.$$

Also:

$$\begin{aligned} \hat{\mathbf{w}}^T\hat{\mathbf{x}}^{(i)} &= \hat{\mathbf{w}}^T\Big(\sum_{j\neq i} a_j \cdot \hat{\mathbf{x}}^{(j)}\Big) \\ &= \hat{\mathbf{w}}^T\Big(\sum_{j\in S} a_j \cdot \hat{\mathbf{x}}^{(j)}\Big) \\ &= \sum_{j\in S} a_j \cdot \hat{\mathbf{w}}^T\hat{\mathbf{x}}^{(j)} \\ &> 0. \end{aligned}$$

So, our assumption is false. Therefore, the VC dimension of the function $f$ in $\mathbb{R}^d$ is at most $d+1$.

Combining (i) and (ii), we can conclude that the VC dimension of the function $f$ in $\mathbb{R}^d$ is $d + 1$.

(b) Prove that $VC(H) \leq \log_2 |H|$ where $H$ is a hypothesis space. (A hypothesis on a set of $n$ points, defines which of two classes can each point belong to. A hypothesis space is a family of all possible hypotheses).

**[Solution]** Suppose a hypothesis space $H$ whose VC-dimension $VC(H) = n$, so there exist $n$ points that $H$ can shatter. We can arbitrarily give 0 or 1 label to each of the points, so there are $2^n$ ways to label them. No matter how the $n$ points are labeled, there exists a hypothesis $h \in H$ which can label them correctly. $H$ must consists at least $2^n$ different hypothesises, that is $|H| \geq 2^n$. So $VC(H) = n \leq \log_2 |H|$.

**Problem 4** (20 pts)
**(Graphical Model / Conditional Independence)**

The conditional independence of two random variables $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$ is often denoted as $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$. Please answer the following questions.

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} := p_{\mathbf{x},\mathbf{y}|\mathbf{z}}(\mathbf{x},\mathbf{y}|\mathbf{z}) = p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})$$

(a) Show that $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$ if and only if the joint distribution for the three variables factors in the following form:

$$p_{\mathbf{x},\mathbf{y},\mathbf{z}} = h(x,z)g(y,z).$$

Need to prove going in both directions and use the fact that $h_1(z) = \int h(x,z)dx$ and $g_1(z) = \int g(y,z)dy$.

**[Solution]**
First, we show that conditional independence implies the desired factorization.

$$\begin{aligned} p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z}) &= p_{\mathbf{z}}(\mathbf{z})p_{\mathbf{x},\mathbf{y}|\mathbf{z}}(\mathbf{x},\mathbf{y}|\mathbf{z}) \\ &= p_{\mathbf{z}}(\mathbf{z})p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) \\ &= h(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{z}) \end{aligned}$$

where we chose $h(\mathbf{x},\mathbf{z}) = p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})$ and $g(\mathbf{y},\mathbf{z}) = p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})p_{\mathbf{z}}(\mathbf{z})$ Now, we show the other direction. For $p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z}) = h(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{z})$, let $h_1(\mathbf{z}) = \sum_{\mathbf{x}} h(\mathbf{x},\mathbf{z})$ and $g_1(\mathbf{z}) = \sum_y g(\mathbf{y},\mathbf{z})$. We need to show that for such $p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z})$, we have $p_{\mathbf{x},\mathbf{y}|\mathbf{z}}(\mathbf{x},\mathbf{y}|\mathbf{z}) = p_{\mathbf{x}|\mathbf{z}}(x|z)p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})$. We first compute

$$\begin{aligned} p_{\mathbf{x},\mathbf{y}|\mathbf{z}}(\mathbf{x},\mathbf{y}|\mathbf{z}) &= \frac{p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z})}{\sum_{\mathbf{x},\mathbf{y}} p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z})} \\ &= \frac{h(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{z})}{\sum_{\mathbf{x},\mathbf{y}} h(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{z})} \\ &= \frac{h(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{z})}{h_1(\mathbf{z})g_1(\mathbf{z})} \end{aligned}$$

Similarly, we can compute $p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = h(\mathbf{x},\mathbf{z})/h_1(\mathbf{z})$ and $p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})g(\mathbf{y},\mathbf{z})/g_1(\mathbf{z})$. This proves the conditional independence, since $p_{\mathbf{x},\mathbf{y}|\mathbf{z}}(\mathbf{x},\mathbf{y}|\mathbf{z}) = p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})$.

(b) Please answer the following questions for the below graphical model, and justify your answer.

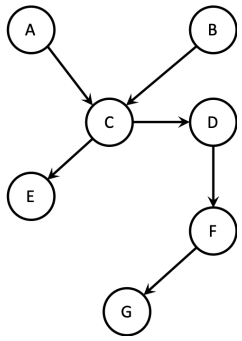


Figure 2: Directed Graphical Model

i Represent the joint distribution of directed graphical model.
**[Solution]** $P(A, B, C, D, E, F, G) = P(A)P(B)P(C|A, B)P(D|C)P(E|C)P(F|D)(G|F)$

ii Are $A$ and $B$ conditionally independent, given $D$ and $F$?
**[Solution]** No, the only path from $A$ to $B$ crosses $C$ with head-to-head relationship. $D$ is a descendent of $C$. Thus given $D$ and $F$, $A$ and $B$ are not conditionally independent.

iii Are $D$ and $E$ conditionally independent, given $C$?
**[Solution]** Yes, the only path from $D$ to $E$ crosses $C$ which is tail-to-tail relationship. Also $C$ is conditioned (blocked node in head-to-head relationship). Thus $D$ and $E$ are conditionally independent given $C$.

iv Are $D$ and $E$ conditionally independent, given $A$ and $B$?
**[Solution]** No, The only path from $D$ to $E$ crosses $C$ (tail-to-tail relationship). From the conditioned variables $A$ and $B$, we cannot know the variable $C$, thus $D$ and $E$ are not conditionally independent given $A$ and $B$.

(c) Represent undirected graph of random variables $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_6, Y_6)\}$ such that the following three constraints are satisfied:

i $Y_i \perp\!\!\!\perp Y_{i+2} | Y_{i+1}$ for $i = 1, \ldots, 4$
ii $X_i \not\perp\!\!\!\perp Y_i$
iii $X_i \perp\!\!\!\perp Y_j | Y_i$ for $j \neq i$

For this graph, express $p(Y_6|X_1, \ldots, X_5)$ in terms of $p(Y_6|Y_5)$ and $p(Y_5|X_1, \ldots, X_5)$.

**[Solution]**

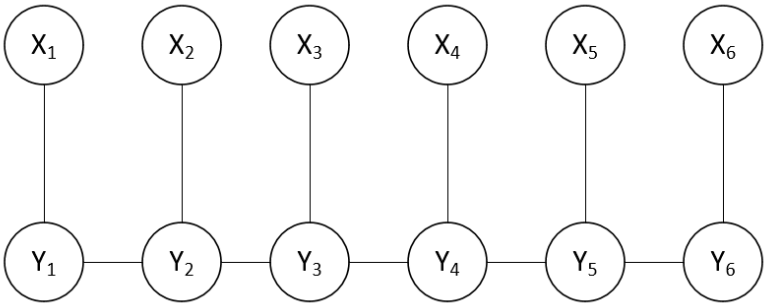


Figure 3: Directed Graphical Model

$$
\begin{aligned}
p(Y_6|X_1, ..., X_5) &= \frac{p(Y_6, X_1, ..., X_5)}{p(X_1, ..., X_5)} = \frac{\sum_{Y_5} p(Y_6, Y_5, X_1, ..., X_5)}{p(X_1, ..., X_5)} \\
&= \frac{\sum_{Y_5} p(X_1, ..., X_5|Y_5)p(Y_6|Y_5)p(Y_5)}{p(X_1, ..., X_5)} \\
&= \frac{\sum_{Y_5} p(Y_5|X_1, ..., X_5)\frac{p(X_1,...,X_5)}{p(Y_5)}p(Y_5)}{p(X_1, ..., X_5)} \\
&= \sum_{Y_5} p(Y_5|X_1, ..., X_5)p(Y_6|Y_5)
\end{aligned}
$$

**Problem 5** (30 pts)
Assume that there are only two grades for qualifying the quality of a apple: high and low grade. In addition, assume that there are only two stores $S_1$ and $S_2$ selling the apples. Each store has different proportion of high-grade (H) and low-grade (L) apples. $S_1$ has $\alpha$% high-grade apples while $S_2$ has $\beta$% high-grade apples. $S_1$ is $\gamma$ times more accessible than S2. Seven apples are bought by a apple lover $O = \{o_1, o_2, ..., o_7\} = \{H, H, H, H, L, L, H\}$ who dies immediately upon purchasing the apples. It is not known where each apple was purchased. No Markov assumption is considered.

(a) What is the probability that the second apple $o_2 = H$ was bought from $q_2 = S_1$, in otherwords, what is $P(q_2 = S_1|o_2 = H, \theta)$. The apples are not necessarily purchased from the same store.

**[Solution]**

From problem, $P(q_i = S_1) = \frac{\gamma}{\gamma+1}$, $P(q_i = S_2) = \frac{1}{\gamma+1}$, $P(o_i = H|q_i = S_1) = \frac{\alpha}{100}$, $P(o_i = H|q_i = S_2) = \frac{\beta}{100}$

$$\begin{aligned} P(q_2 = S_1|o_2 = H, \theta) &= \frac{P(q_2 = S_1, o_2 = H, \theta)}{P(o_2 = H, \theta)} = \frac{P(o_2 = H|q_2 = S_1, \theta)P(q_2 = S_1)}{\sum_{i=1}^{2} P(q_2 = S_i, o_2 = H, \theta)} \\ &= \frac{\frac{\gamma}{\gamma+1}\frac{\alpha}{100}}{\frac{\gamma}{\gamma+1}\frac{\alpha}{100} + \frac{1}{\gamma+1}\frac{\beta}{100}} = \frac{\gamma\alpha}{\gamma\alpha + \beta} \end{aligned}$$

(b) Define the auxiliary function $Q(\theta, \bar{\theta})$ where $\theta = \{\alpha, \beta, \gamma\}$.

**[Solution]** The auxiliary function is defined as

$$Q(\theta, \bar{\theta}) = \mathbb{E}_{P(H|D,\theta)}[\log P(H, D|\bar{\theta})]$$

Substitute $\lambda_i = P(q_i = S_1|o_i, \theta)$,

$$\begin{aligned} \lambda_i &= \frac{P(o_i|q_i = S_1, \theta)P(q_i = S_1|\theta)}{\sum_j P(o_i|q_i = S_j, \theta)P(q_i = S_j|\theta)} \\ &= \frac{\frac{\gamma}{\gamma+1}\alpha^{o_i}(100-\alpha)^{1-o_i}}{\frac{\gamma}{\gamma+1}\alpha^{o_i}(100-\alpha)^{1-o_i} + \frac{1}{\gamma+1}\beta^{o_i}(100-\beta)^{1-o_i}} \\ &= \frac{\gamma\alpha^{o_i}(100-\alpha)^{1-o_i}}{\gamma\alpha^{o_i}(100-\alpha)^{1-o_i} + \beta^{o_i}(100-\beta)^{1-o_i}} \end{aligned}$$

Therefore,

$$Q(\theta, \bar{\theta}) = \sum_i \left[ \lambda_i \log[\frac{\bar{\gamma}}{\bar{\gamma}+1}\bar{\alpha}^{o_i}(100-\bar{\alpha})^{1-o_i}] + (1-\lambda_i)\log[\frac{1}{\bar{\gamma}+1}\bar{\beta}^{o_i}(100-\bar{\beta})^{1-o_i}] \right]$$

(c) Derive the M-step to find $\bar{\theta}$ that maximizes the auxiliary function.

**[Solution]** M step is to find $\bar{\theta}$ that maximize the auxiliary function,

$$\bar{\theta} = \arg\max_{\bar{\theta}} Q(\theta, \bar{\theta})$$

To calculate $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$, differentiate $Q$ w.r.t. $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$ and find value that makes $\frac{\partial Q}{\partial \bar{\theta}} = 0$.

$$\begin{aligned} \bar{\alpha} &= \frac{\sum_i \lambda_i o_i}{\sum_i \lambda_i} \\ \bar{\beta} &= \frac{\sum_i (1-\lambda_i) o_i}{\sum_i (1-\lambda_i)} \\ \bar{\gamma} &= \frac{\sum_i \lambda_i}{\sum_i (1-\lambda_i)} \end{aligned}$$