Recitation 8
Graphical Models

10-601: Introduction to Machine Learning

4/23/2021

1 Warm-Up: Probability Review

$\mathbf{X_1}$	$\mathbf{X_2}$	Probability
0	0	0.45
1	0	0.15
0	1	0.2
1	1	0.2

Table 1: Joint Probability Table

1. What is the joint probability $P(X_1 = 1, X_2 = 0)$? From table, 0.15.
2. What is the marginal probability $P(X_2 = 0)$? Add the table items, 0.45 + 0.15 = 0.6.

3. What is the conditional probability $P(X_1 = 1|X_2 = 0)$?
   Definition of conditional probability:

$$\frac{P(X_1 = 1, X_2 = 0)}{P(X_2 = 0)} = \frac{0.15}{0.6} = 0.25$$

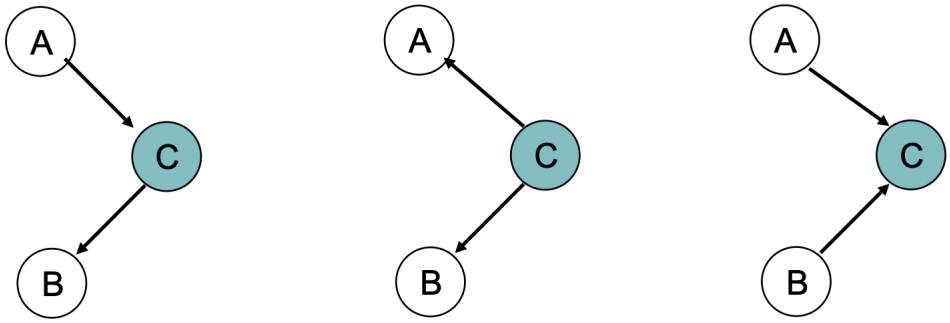

Figure 1: A and B are d-seperated by C in the first two graphs, but not in the third.

2 Bayesian Networks

2.1 D-seperation

Suppose we have three sets of random variables (X, Y, and Z). X and Y are **d-seperated** by Z (and therefore conditionally independent given Z) if and only if every path from from every variable in X to every variable in Y is **blocked**. A path is blocked if either:

- Arrows meet head-to-tail or tail-to-tail at a node in Z
- Arrows on the path meet head-to-head at a node, and neither that node, nor any of its descendants, is in Z

2.2 Practice problems

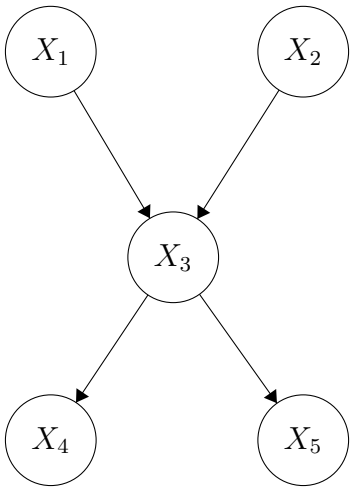

Figure 2: Graphical Model

1. Write down the factorization of the above directed graphical model.

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_3)P(X_5|X_3)$$

2. Given X_3, what are the relationships (cond. independent or not) between the random variables listed below
   - $(X_1$_____$X_4)|X_3$
   - $(X_1$_____$X_2)|X_3$
   - $(X_4$_____$X_5)|X_3$

   1. conditionally independent 2. not conditionally independent 3. conditionally independent

3. Given the graph structure and assuming all variables are boolean valued, how many parameters are required to learn the graphical model?

   $1 + 1 + 4 + 2 + 2 = 10$

4. Without the Bayesian network, how many parameters are required to learn the joint probability model of these five random variables?

   $2^5 - 1 = 31$

3 Expectation Maximization

Consider the following problem set-up. We have two coins, coin A and coin B. Each coin has a probability of coming up heads that we would like to estimate, i.e. θ_A, θ_B. Suppose we now have the following procedure:

- Randomly choose one of the coins.
- Using the chosen coin, perform 6 independent coin flips.
- Repeat the process 3 times.

Now suppose we observe the following data for each of the 3 trials:

Trial 1	H	H	H	H	H	T
Trial 2	T	T	T	T	T	H
Trial 3	H	T	T	T	T	H

Table 2: Coin Flip Data

If we knew what coin was selected for each trial, we could easily use MLE estimation to estimate $\hat{\theta}_A, \hat{\theta}_B$. Instead, we will assume that we don't know which coin was selected, and will have to learn the latent factor describing which coin was selected to maximize the likelihood.

We would like to use the EM algorithm to estimate $\hat{\theta}_A, \hat{\theta}_B$. Rather than picking the single most likely completion of the missing coin assignments on each iteration, the expectation maximization algorithm computes probabilities for each possible completion of the missing data, using the current parameters. These probabilities are used to create a weighted training set consisting of all possible completions of the data.

To estimate the unknown values in the E-step at iteration t, for each trial i we will compute weights α_i, β_i representing how likely the data is to come from coin A versus coin B.

As a results,

$$\alpha_i = \frac{P(D_i|\hat{\theta}_A^{(t)})}{P(D_i|\hat{\theta}_A^{(t)}) + P(D_i|\hat{\theta}_B^{(t)})}$$

and,

$$\beta_i = \frac{P(D_i|\hat{\theta}_B^{(t)})}{P(D_i|\hat{\theta}_A^{(t)}) + P(D_i|\hat{\theta}_B)^{(t)}} = 1 - \alpha_i$$

where D_i is the data for trial i.

Once we have estimated α_i, β_i for each trial, we will then perform the M-step of maximizing the parameters $\hat{\theta}_A, \hat{\theta}_B$.

To do so, we will compute

$$\hat{\theta}_A^{(t+1)} = \sum_i \frac{\alpha_i * \text{Number of heads in trial } i}{\alpha_i * \text{Number of heads in trial } i + \alpha_i * \text{Number of tails in trial } i}$$

$$\hat{\theta}_B^{(t+1)} = \sum_i \frac{\beta_i * \text{Number of heads in trial } i}{\beta_i * \text{Number of heads in trial } i + \beta_i * \text{Number of tails in trial } i}$$

Essentially, we are performing an MLE estimate weighted by our best guess of which coin was selected for each trial under our current parameters.

Suppose our initial guesses are $\hat{\theta}_A^{(0)} = 0.6, \hat{\theta}_B^{(0)} = 0.5$

1. Compute α_i, β_i to 2 decimal places for each trial i.

   Trial 1: $P(D_1|\hat{\theta}_A^{(0)}) = 0.6^5 * 0.4$
   $P(D_1|\hat{\theta}_B^{(0)}) = 0.5^5 * 0.5$
   $\alpha_1 = \frac{0.6^5 * 0.4}{0.6^5 * 0.4 + 0.5^5 * 0.5} \approx 0.67$
   $\beta_1 \approx 0.33$

   Trial 2: $P(D_1|\hat{\theta}_A^{(0)}) = 0.6^1 * 0.4^5$
   $P(D_1|\hat{\theta}_B^{(0)}) = 0.5^1 * 0.5^5$
   $\alpha_2 = \frac{0.6^1 * 0.4^5}{0.6^1 * 0.4^5 + 0.5^3 * 0.5^3} \approx 0.28$
   $\beta_2 \approx 0.72$

   Trial 3: $P(D_1|\hat{\theta}_A^{(0)}) = 0.6^2 * 0.4^4$
   $P(D_1|\hat{\theta}_B^{(0)}) = 0.5^2 * 0.5^4$
   $\alpha_2 = \frac{0.6^2 * 0.4^4}{0.6^2 * 0.4^4 + 0.5^2 * 0.5^4} \approx 0.37$
   $\beta_2 \approx 0.63$

2. With the found values of α_i, β_i, compute $\hat{\theta}_A^{(1)}, \hat{\theta}_B^{(1)}$ in the M-step

$$\hat{\theta}_A^{(t+1)} = \sum_i \frac{\alpha_i * \text{Number of heads in trial } i}{\alpha_i * \text{Number of heads in trial } i + \alpha_i * \text{Number of tails in trial } i}$$

$$\hat{\theta}_A^{(t+1)} = \frac{0.67 * 5 + 0.28 * 1 + 0.37 * 2}{0.67 * 5 + 0.28 * 1 + 0.37 * 2 + 0.67 * 1 + 0.28 * 5 + 0.37 * 4} = 0.552$$

$$\hat{\theta}_B^{(t+1)} = \sum_i \frac{\beta_i * \text{Number of heads in trial } i}{\beta_i * \text{Number of heads in trial } i + \beta_i * \text{Number of tails in trial } i}$$

$$\hat{\theta}_B^{(t+1)} = \frac{0.33 * 5 + 0.72 * 1 + 0.63 * 2}{0.33 * 5 + 0.72 * 1 + 0.63 * 2 + 0.33 * 1 + 0.72 * 5 + 0.63 * 4} = 0.360$$

4 Gaussian Naive Bayes

Priors

$$P(Y=1) = \frac{\#D\{Y=1\}}{|D|}$$

$$P(Y=0) = \frac{\#D\{Y=0\}}{|D|}$$

Class conditional means, variances

$$\hat{\mu}_{m,k} = \frac{1}{\sum_{i=1}^{N} I(Y^{(i)}=k)} \sum_{i=1}^{N} \mathbf{x}_m^{(i)} * I(Y^{(i)}=k)$$

$$\hat{\sigma}_{m,k}^2 = \frac{1}{\sum_{i=1}^{N} I(Y^{(i)}=k)} \sum_{i=1}^{N} (\mathbf{x}_m^{(i)} - \hat{\mu}_{m,k})^2 * I(Y^{(i)}=k)$$

1. Write down the Naive Bayes assumption.

   Naive Bayes assumes that features are conditionally independent given the class, and weights each feature equally.

2. When would we use Gaussian Naive Bayes?

   When X_i is continuous, we may use Gaussian Naive Bayes to assume that $P(Xi|Y = yk)$ follows a normal distribution.

3. Consider a simple Gaussian Naive Bayes example. Let's say we want to use predict if a person will go hiking today based on the temperature outside.

Temperature (F)	Hiking?
50.8	No
67.6	Yes
39.7	No
36.5	No
72.4	Yes

   (a) Step 1: Find the class priors.

$$P(Y=Yes) = \frac{\#D\{Y=Yes\}}{|D|} = 2/5$$

$$P(Y=No) = \frac{\#D\{Y=No\}}{|D|} = 3/5$$

(b) Step 2: Find the class conditional mean, variance for X (temperature).

For each class k, find $\hat{\mu}_{temp,k}$, the mean over all values for the temperature where the class is k.

$$\hat{\mu}_{temp,Yes} = \frac{1}{\sum_{i=1}^{5} I(Y^{(i)} = Yes)} \sum_{i=1}^{5} \mathbf{x}_{temp}^{(i)} * I(Y^{(i)} = Yes)$$

$$= \frac{1}{2}(50.8 * 0 + 67.6 * 1 + 39.7 * 0 + 36.5 * 0 + 72.4 * 1) = 70.0$$

$$\hat{\mu}_{temp,No} = \frac{1}{\sum_{i=1}^{5} I(Y^{(i)} = No)} \sum_{i=1}^{5} \mathbf{x}_{temp}^{(i)} * I(Y^{(i)} = No)$$

$$= \frac{1}{3}(50.8 * 1 + 67.6 * 0 + 39.7 * 1 + 36.5 * 1 + 72.4 * 0) = 42.33$$

Similarly, for each class k, calculate the variance.

$$\hat{\sigma}^2_{temp,Yes} = \frac{1}{\sum_{i=1}^{5} I(Y^{(i)} = Yes)} \sum_{i=1}^{5} (\mathbf{x}_{temp}^{(i)} - \hat{\mu}_{temp,Yes})^2 * I(Y^{(i)} = Yes)$$

$$\frac{1}{2}((67.6 - 70)^2 + (72.4 - 70)^2)) = 5.76$$

Thus, the standard deviation σ is $\sqrt{5.76} = 2.4$

$$\hat{\sigma}^2_{temp,No} = \frac{1}{\sum_{i=1}^{5} I(Y^{(i)} = No)} \sum_{i=1}^{5} (\mathbf{x}_{temp}^{(i)} - \hat{\mu}_{temp,No})^2 * I(Y^{(i)} = No)$$

$$\frac{1}{3}((50.8 - 42.33)^2 + (39.7 - 42.33)^2) + (36.5 - 42.33)^2) = 37.55$$

Thus, the standard deviation σ is $\sqrt{37.55} = 6.128$

(c) Calculate the following class-conditional probabilities for Temperature = 65 :

$P(X_{temp} = 69.5|Y = k)$ We estimate the class-conditional probability as a Gaussian distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

For class Y=Yes:

$$P(X_{temp} = 69.5|Y = Yes) = \frac{1}{\sqrt{2\pi}\sigma_{temp,Yes}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_{temp,Yes}}{\sigma_{temp,Yes}}\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi} * 2.4} \exp\left(-\frac{1}{2}\left(\frac{69.5 - 70.0}{2.4}\right)^2\right) = 0.1627$$

For class Y=No:

$$P(X_{temp} = 69.5|Y = No) = \frac{1}{\sqrt{2\pi}\sigma_{temp,No}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_{temp,No}}{\sigma_{temp,No}}\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi} * 6.128} \exp\left(-\frac{1}{2}\left(\frac{69.5 - 42.33}{6.128}\right)^2\right) = 5.46e - 19$$

(d) The red graph represents the class conditional mean for Y=Yes, and the green graph represents the class conditional mean for Y=No. Based on your results from the last question, interpret the graph below and explain how the output predictions $P(Y = k|X_{temp})$ are affected by different feature values.

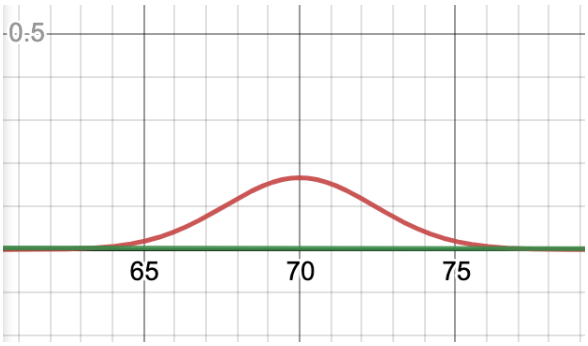

Figure 3: Class conditional probability for temperature, Y=Yes

Consistent with the answer from part c), we can see that at a temperature of 69.5 Fahrenheit the class-conditional probability for class "Yes" (the red graph) is around 0.1627, while for class "No" it is around 0.

Within this range (temperature is between 65 and 75), "Yes" will always be predicted as an output.