STAT 479: Homework 2

Due: 11:59PM Feb 11, 2025 by Canvas

Part 1: Naive Bayes

1. **MLE for Gaussian Naive Bayes** (20 points)

Naive Bayes assumes that features $X = (X_1, X_2, \ldots, X_d)$ are conditionally independent given the class Y, and that $X_i|Y \sim \mathcal{N}(\mu_{i,Y}, \sigma^2_{i,Y})$.

- Features are conditionally independent given the class:

$$P(X|Y) = \prod_{i=1}^{d} P(X_i|Y).$$

- Each feature $X_i|Y = k$ follows a Gaussian distribution:

$$P(X_i|Y = k) = \frac{1}{\sqrt{2\pi\sigma^2_{i,k}}} \exp\left(-\frac{(X_i - \mu_{i,k})^2}{2\sigma^2_{i,k}}\right).$$

(a) **Likelihood Function**: Write the likelihood of the observed data $\{X_{i,j}\}_{j:Y_j=k}$, assuming $X_i|Y = k$ is Gaussian. Select the correct expression:

A. $\prod_{j:Y_j=k} \prod_{i=1}^{d} P(X_{i,j}|Y = k)$

B. $\prod_{j:Y_j=k} \prod_{i=1}^{d} \frac{1}{2\pi\sigma^2_{i,k}} \exp\left(-\frac{(X_{i,j}-\mu_{i,k})^2}{\sigma^2_{i,k}}\right)$

C. $\prod_{j:Y_j=k} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2_{i,k}}} \exp\left(-\frac{(X_{i,j}-\mu_{i,k})^2}{2\sigma^2_{i,k}}\right)$

D. $\prod_{j:Y_j=k} \prod_{i=1}^{d} P(Y = k) \cdot P(X_{i,j})$

(b) **Log-Likelihood**: Write the log-likelihood for $\mu_{i,k}$ and $\sigma^2_{i,k}$. For notation convenience, let N_k be the number of samples with label k: $N_k = \sum_{j=1}^{n} \mathbb{I}(Y_j = k)$. Select the correct expression:

A. $-\frac{N_k}{2} \log(2\pi\sigma^2_{i,k}) - \frac{1}{2\sigma^2_{i,k}} \sum_{j:Y_j=k} (X_{i,j} - \mu_{i,k})^2$

B. $-\frac{1}{2} \log(2\pi\sigma^2_{i,k}) - \frac{1}{\sigma^2_{i,k}} \sum_{j:Y_j=k} (X_{i,j} - \mu_{i,k})^2$

C. $\sum_{j:Y_j=k} \log P(X_{i,j}|Y = k) + \log P(Y = k)$

D. $-\frac{N_k}{2} \log(2\pi) + \frac{\sum_{j:Y_j=k}(X_{i,j}-\mu_{i,k})^2}{2\sigma^2_{i,k}}$

(c) **MLE for** $\mu_{i,k}$: Derive the MLE for $\mu_{i,k}$. Select the correct estimator:

A. $\widehat{\mu_{i,k}} = \frac{\sum_{j:Y_j=k} X_{i,j}}{N_k}$

B. $\widehat{\mu_{i,k}} = \frac{\sum_{j=1}^{n} X_{i,j}}{n}$

C. $\widehat{\mu_{i,k}} = \frac{\sum_{j:Y_j=k} X_{i,j}^2}{N_k}$

D. $\widehat{\mu_{i,k}} = \sum_{j:Y_j=k} X_{i,j}$

Answer:

(a)

(b)

(c)

2. **Modified Naive Bayes with Feature Dependencies** (20 points)

Background: In the standard Naive Bayes classifier, we assume that all features $X_1, X_2, \ldots, X_d$ are conditionally independent given the class label Y:

$$P(X_1, X_2, \ldots, X_d \mid Y) = \prod_{i=1}^{d} P(X_i \mid Y)$$

However, in many real-world scenarios, features exhibit dependencies. To address this, we modify the Naive Bayes model to account for such dependencies. Specifically, we introduce a structure where some features are conditionally dependent on others.

Problem Statement: Consider a dataset with four features X_1, X_2, X_3, X_4 and a binary class label $Y \in \{0, 1\}$. Instead of assuming full independence, we impose the dependency structure captured in the following joint distribution:

$$P(X_1, X_2, X_3, X_4 \mid Y) = P(X_1 \mid Y)P(X_2, X_3 \mid X_1, Y)P(X_4 \mid Y)$$

Here, $P(X_2, X_3 \mid X_1, Y)$ is modeled as a joint Gaussian distribution with mean and covariance matrix dependent on Y.

Given the probability distributions:

- $P(Y = 0) = 0.5, P(Y = 1) = 0.5$
- $P(X_1 \mid Y)$ and $P(X_4 \mid Y)$ are **standard normal** univariate Gaussians.
- $P(X_2, X_3 \mid X_1, Y)$ follows a multivariate Gaussian distribution:

- **For** $Y = 0$:

$$\mu_{X_2,X_3} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \alpha X_1, \quad \Sigma_{X_2,X_3} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- **For** $Y = 1$:

$$\mu_{X_2,X_3} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \alpha X_1, \quad \Sigma_{X_2,X_3} = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1.5 \end{bmatrix}$$

(a) **Computing** $P(X \mid Y)$
What is $P(X_1 = 0, X_2 = 2, X_3 = 3, X_4 = 0 \mid Y = 0)$ under this modified Naive Bayes model?

A. 0.001
B. 0.01
C. 0.03
D. 0.3

(b) **Computing** $P(X, Y)$
What is $P(X_1 = 0, X_2 = 2, X_3 = 3, X_4 = 0, Y = 0)$ under this modified Naive Bayes model?

A. 0.001
B. 0.015
C. 0.2
D. 0.5

(c) **Model Complexity** How many parameters are required to fully parameterize this modified Naive Bayes model compared to the standard Naive Bayes model?

A. **More than standard Naive Bayes** – because the dependency structure introduces additional covariance terms in $P(X_2, X_3 \mid X_1, Y)$.
B. **Fewer than standard Naive Bayes** – because the dependency structure reduces the total number of independent conditional distributions.
C. **The same as standard Naive Bayes** – because each feature still depends on the class label Y.
D. **It depends on the dataset** – the number of parameters cannot be determined without knowing the data distribution.

Answer:

(a)

(b)

(c)

Part 2: Bayesian Networks

3. **Conditional Independence and D-Separation** (20 points)

Consider the following Bayesian Network, where A, B, C, and D are random variables:

$$A \to B \to D, \quad A \to C \to D.$$

(a) **Markov Property**: Which of the following statements about conditional independence in this Bayesian Network is correct?

A. A and D are conditionally independent given B.
B. A and D are conditionally independent given C.
C. B and C are conditionally independent given D.
D. B and C are conditionally independent given A.

(b) **Joint Distribution**: Which of the following correctly represents the joint probability $P(A, B, C, D)$?

A. $P(A, B, C, D) = P(A)P(B|A)P(C|A)P(D|B, C)$
B. $P(A, B, C, D) = P(A)P(B)P(C|A)P(D|B, C)$
C. $P(A, B, C, D) = P(A)P(B|A)P(C)P(D|B, C)$
D. $P(A, B, C, D) = P(A|B)P(B|C)P(C|D)P(D)$

(c) **d-separation**: Which of the following pairs of variables are d-separated in the given network, assuming no evidence is observed?

A. A and D
B. B and C
C. A and C
D. None of the above
E. All of the above

(d) **d-separation**: Which of the following pairs of variables are d-separated in the given network, assuming B is observed?

A. A and D
B. B and C
C. A and C
D. None of the above
E. All of the above

(e) **D-Separation and Deep Generative Models:** Suppose a deep generative model learns latent variables Z that mediate dependencies between observed variables. Which of the following statements is true?

(a) If Z explains the correlation between X and Y, then conditioning on Z should make X and Y independent.

(b) Adding a latent variable always increases dependencies between observed variables.

(c) If Z is a common parent of X and Y, then X and Y are always independent.

(d) If Z is observed, it has no effect on the conditional independence structure of X and Y.

Answer:

(a)

(b)

(c)

(d)

(e)

Part 3: HMM

4. **HMM Theory** (20 points)

Recall our discussion of *Hidden Markov Models (HMMs)*, which are statistical models where an **unobserved (hidden) state sequence** influences an **observed sequence** of data.

(a) **Long-Term Behavior:** If an HMM runs for many timesteps, we expect the probabilities of being in each state to:

A. Continue changing randomly with no pattern.

B. Settle into stable values over time.

C. Eventually become equal for all states.

D. Always remain the same as the initial probabilities.

(b) **Interpretation of $(P^n)_{ij}$:** Let P be the matrix of transition probabilities. The value of $(P^n)_{ij}$ in the transition matrix raised to the power n represents:

A. The probability of being in state j after n steps, regardless of the initial state.

B. The probability of transitioning from state i to state j in exactly n steps.

C. The long-run proportion of time spent in state j.

D. The expected number of transitions from i to j over n steps.

Example: Consider the transition matrix:

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Then P^2 gives:

$$P^2 = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \times \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix}$$

where $(P^2)_{12} = 0.39$.

(c) **Powers of the Transition Matrix:** The matrix P^n, representing the transition probabilities after n steps, has the following property:

A. It becomes a diagonal matrix as n increases.

B. It converges to a matrix where all rows are identical.

C. It remains the same as P for all n.

D. It fluctuates indefinitely without a pattern.

Hint: In many Markov chains, as n grows, the transition probabilities stabilize, meaning that each row of P^n approaches a unique limiting distribution (the stationary distribution). This suggests that long-run behavior is independent of the initial state.

Example: For the transition matrix P above, as $n \to \infty$, P^n converges to:

$$P^\infty \approx \begin{bmatrix} 0.57 & 0.43 \\ 0.57 & 0.43 \end{bmatrix}$$

Answer:

(a)

(b)

(c)

5. **The Forward-Backward Algorithm** (20 points)

A key question in HMMs is how to compute the probability of a hidden state at a particular time step given all observations.

To do this efficiently, we use the **Forward-Backward Algorithm**, which introduces two key quantities:

- The **forward probability**:

$$\alpha_t(Z_t) = P(X_1, \ldots, X_t, Z_t)$$

which represents the probability of the first t observations and the current state.

- The **backward probability**:

$$\beta_t(Z_t) = P(X_{t+1}, \ldots, X_n | Z_t)$$

which represents the probability of future observations given the current state.

Using these definitions, answer the following:

(a) **Recursive Formula for** $\alpha_t(Z_t)$ The forward probability $\alpha_t(Z_t)$ can be computed recursively using the previous timestep $\alpha_{t-1}(Z_{t-1})$:

A. $\alpha_t(Z_t) = P(X_t|Z_t) \sum_{Z_{t-1}} P(Z_t|Z_{t-1})\alpha_{t-1}(Z_{t-1})$
B. $\alpha_t(Z_t) = P(X_t|Z_t)P(Z_t)$
C. $\alpha_t(Z_t) = \sum_{Z_{t-1}} P(X_t|Z_t)P(Z_t|Z_{t-1})\alpha_{t-1}(Z_{t-1})\beta_{t+1}(Z_t)$
D. $\alpha_t(Z_t) = P(X_t|Z_t)P(Z_t|Z_{t-1})$

(b) **Recursive Formula for $\beta_t(Z_t)$** Similarly, the backward probability $\beta_t(Z_t)$ can be computed recursively from the next timestep $\beta_{t+1}(Z_{t+1})$:

A. $\beta_t(Z_t) = \sum_{Z_{t+1}} P(Z_{t+1}|Z_t)P(X_{t+1}|Z_{t+1})\beta_{t+1}(Z_{t+1})$
B. $\beta_t(Z_t) = P(X_{t+1}|Z_t)P(Z_{t+1}|Z_t)\beta_{t+1}(Z_{t+1})$
C. $\beta_t(Z_t) = \sum_{Z_{t+1}} P(Z_t|Z_{t+1})P(X_{t+1}|Z_t)\beta_{t+1}(Z_{t+1})$
D. $\beta_t(Z_t) = P(Z_{t+1}|Z_t)P(X_t|Z_t)\beta_t(Z_t)$

(c) **Computing $P(Z_t|X)$** Now, using $\alpha_t(Z_t)$ and $\beta_t(Z_t)$, the posterior probability of Z_t given the full sequence X is:

A. $P(Z_t|X) = \frac{\alpha_t(Z_t)\beta_t(Z_t)}{P(X)}$
B. $P(Z_t|X) = \frac{P(X|Z_t)P(Z_t)}{P(X)}$
C. $P(Z_t|X) = \frac{P(X_t|Z_t)P(Z_t|X_{1:t-1})}{P(X_t|X_{1:t-1})}$
D. $P(Z_t|X) = \frac{\alpha_t(Z_t)P(X)}{\beta_t(Z_t)}$

(d) **Why is the Forward-Backward Algorithm Efficient?** The Forward-Backward algorithm provides an efficient way to compute $P(Z_t|X)$ for all time steps t. Which of the following best explains how it reduces computation compared to a naive approach?

A. Instead of summing over all possible hidden state sequences, it breaks the problem into two recursive computations—one moving forward in time, and one moving backward—reducing the complexity from $O(n \cdot |Z|^n)$ to $O(n|Z|^2)$.
B. It replaces probabilities with log-probabilities, converting multiplication into addition, which reduces computational cost.
C. It precomputes $P(X)$, avoiding the need for normalization in the final computation of $P(Z_t|X)$.
D. It finds the most probable sequence of hidden states using dynamic programming, rather than computing marginal probabilities for each state individually.

Answer:

(a)

(b)

(c)

(d)