CS 4824/ECE 4424:
Gaussian Naïve Bayes

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Naïve Bayes in a Nutshell

- Bayes Rule:
  \[ P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n | Y = y_j)} \]

- Assuming conditional independence among \( X_i \)'s
  \[ P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)\prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i | Y = y_j)} \]

- How to pick the most probable \( Y \) for \( X^{\text{New}} = <X_1, X_2, \ldots, X_n> \)?

\[ Y^{\text{New}} \leftarrow \arg \max_{y_k} P(Y = y_k)\prod_i P(X_i^{\text{New}} | Y = y_k) \]
Naïve Bayes algorithm - discrete $X_i$

- Train Naïve Bayes (examples)
  - For each value $y_k$
    - Estimate $\pi_k = P(Y = y_k)$
    - For each value $x_{ij}$ of each attribute $X_i$
      - Estimate $\theta_{ijk} = P(X = x_{ij} | Y = y_k)$
    - Note: Prob. must sum to 1 so we only need to estimate $n-1$ of these

- Classify $X^{\text{New}}$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{New}} | Y = y_k)$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk}$
Another way to view Naïve Bayes (boolean $Y$)

Decision rule:

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

$$1 \geq \frac{P(Y=1 | X_1 \ldots X_n)}{P(Y=0 | X_1 \ldots X_n)} = \frac{P(Y=1) \prod \limits_{i} P(X_i | Y=1)}{P(Y=0) \prod \limits_{i} P(X_i | Y=0)}$$

$$0 \geq \ln \frac{P(Y=1) \prod \limits_{i} P(X_i | Y=1)}{P(Y=0) \prod \limits_{i} P(X_i | Y=0)}$$

$$0 \geq \ln \frac{P(Y=1)}{P(Y=0)} + \sum \limits_{i} \ln \frac{P(X_i | Y=1)}{P(X_i | Y=0)}$$

Linear sum of a prior term and conditional prob. term.
What if we have continuous $X_i$

- For example, image classification
  - $X_i$ is the $i$th pixel, $Y =$ mental state

- We still have

$$P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- How to represent $P(X_i | Y)$?
What if we have continuous $X_i$

- For example, image classification
  - $X_i$ is the $i$th pixel, $Y =$ mental state
  
- Gaussian Naïve Bayes (GNB) assumes
  
  \[
  P(X_i | Y) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left( \frac{x - \mu_{ik}}{\sigma_{ik}} \right)^2 \right)
  \]

- Sometimes assume $\sigma_{ik}$
  - is independent of $Y$ (i.e., $\sigma_i$)
  - is independent of $X_i$ (i.e., $\sigma_k$)
  - or both (i.e., $\sigma$)

$k$ is the class label

$i$ is the feature

$\var \text{ means}$

what are the implications of these assumptions?
Gaussian Naïve Bayes algorithm: continuous $X_i$ but discrete $Y$

- **Train Naïve Bayes (examples)**
  - For each value $y_k$
    - Estimate $\pi_k = P(Y = y_k)$
    - For each value $x_{ij}$ of each attribute $X_i$
      - Estimate class conditional $\mu_{ik}$ and variance $\sigma_{ik}$
    - Note: Prob. must sum to 1 so we only need to estimate $n-1$ of these

- **Classify $X^{\text{New}}$**
  - $Y^{\text{New}} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X^{\text{New}}_i | Y = y_k)$
  - $Y^{\text{New}} \leftarrow \arg\max_{y_k} \pi_k \prod_i \mathcal{N}(X^{\text{New}}_i, \mu_{ik}, \sigma_{ik})$
Estimating parameters: continuous $X_i$ but discrete $Y$

- MLE
  
  \[ \hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y_j = y_k)} \sum_j X_i^j \delta(Y_j = y_k) \]
  
  \[ \sigma_{ik}^2 = \frac{1}{\sum_j \delta(Y_j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y_j = y_k) \]
Gaussian Naïve Bayes - decision surface

- Assume \( Y = \text{PlayBasketball} \) (boolean) \( X_1 = \text{Height} \) \( X_2 = \text{Age} \).
- \( Y^{\text{New}} \leftarrow \arg \max_{y_k} P(Y \mid y_k) \prod_i P(X_i^{\text{New}} \mid Y = y_k) \); assume \( P(Y=1) = 0.5 \).
What is the minimum possible error?

- Best case:
  - Conditional independence assumption is satisfied
  - We can perfectly estimate $P(Y)$, $P(X|Y)$ (i.e. infinite training data)
But...

- Naïve Bayes allows estimating $P(Y|X)$ by learning $P(Y)$ and $P(X|Y)$
- Why not learn $P(Y|X)$ directly?