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:
  \[
P(Y = y_k \mid X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n \mid Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n \mid Y = y_j)}
  \]
- Assuming conditional independence among \(X_i\)'s
  \[
P(Y = y_k \mid X_1, \ldots, X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}
  \]
- How to pick the most probable \(Y\) for \(X^{\text{New}} = <X_1, X_2, \ldots, X_n>\)?
  \[
  Y^{\text{New}} \leftarrow \arg\ max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{New}} \mid Y = y_k)
  \]
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 $X^{New}$**
  - $Y^{New} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_{i} P(X^{New}_i | Y = y_k)$
  - $Y^{New} \leftarrow \arg \max_{y_k} \pi_k \prod_{i} \theta_{ijk}$
Another way to view Naïve Bayes (boolean Y)

- Decision rule:
What if we have continuous $X_i$

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

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

- 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, $Y = \text{mental state}$

- Gaussian Naïve Bayes (GNB) assumes
  
  $$P(X_i \mid Y) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu_{ik}}{\sigma_{ik}} \right)^2}$$

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

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

- Assume $Y=$PlayBasketball (boolean) $X_1=$Height $X_2=$Age
  - $Y_{\text{New}} \leftarrow \text{arg max}_{y_k} P(Y | y_k) \prod_{i} P(X_{i}^{\text{New}} | Y = y_k)$; 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 $P(Y \mid X)$ by learning $P(Y)$ and $P(X \mid Y)$

- Why not learn $P(Y \mid X)$ directly?