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 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 $X^{\text{New}}$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_{i} P(X_i^{\text{New}} | Y = y_k)$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} \pi_k \prod_{i} \theta_{ijk}$
How to estimate parameters: discrete-valued $Y, X_i$

- Maximum likelihood estimates (MLE’s)

$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$

$\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} \mid Y = y_k) = \frac{\#D\{X_j = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$
Naïve Bayes issue #1

- Often $X_i$’s are not really conditionally independent
  - We can still use Naïve Bayes and works “pretty well”
  - Often results in right classification but not right prob.

- What is the effect on estimated $P(Y|X)$?
  - Extreme case: what if we have two copies $X_i=X_k$
    - $P(Y=1|X) = P(Y=1) P(X_1|Y=1) P(X_2|Y=1) ... P(X_i|Y=1) ... P(X_k|Y=1)$
Naïve Bayes issue #2

- If unlucky, the MLE estimate for $P(X_i \mid Y)$ might be zero
  - Why worry about just one parameter?

- What can we do to address it?
Using MAP estimation: discrete-valued $Y, X_i$

- Maximum a posteriori estimate (MAP)
  - What should be our prior?
  - How to incorporate the prior into the MLE?

\[
\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}
\]

\[
\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} \mid Y = y_k) = \frac{\#D\{X_j = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}
\]
Using MAP estimation: discrete-valued $Y, X_i$

- Maximum a posteriori estimate (MAP)
  - (Beta, Dirichlet prior)

\[
\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}
\]

\[
\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} | Y = y_k) = \frac{\#D\{X_j = x_{ij} \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}
\]
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

![](image.png)

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

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

- **MLE**
  - $\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X^j_i \delta(Y^j = y_k)$
  - $\sigma^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$