CS 4824/ECE 4424: Generative vs. Discriminative Classifiers

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Generative vs. discriminative classifiers

- Training classifiers involve estimating $f: X \rightarrow Y$ or $P(Y | X)$

- Generative classifiers (e.g., Naïve Bayes)
  - Assumes some functional form for $P(X | Y)$, $P(Y)$
  - Estimates parameters of $P(X | Y)$, $P(Y)$ from training data
  - Use Bayes rule to calculate $P(Y | X)$
  - $Y$ is boolean

- Discriminative classifiers (e.g., Logistic Regression)
  - Assumes some functional form for $P(Y | X)$
  - Estimates parameters of $P(Y | X)$ directly from training data

**NOTE**: Even through our derivation of the form of $P(Y | X)$ made GNB-style assumptions, the *training procedure* for logistic regression does not!
Use Naïve Bayes or Logistic Regression?

- Consider
  - Restrictiveness of modeling assumption
    - How well we can learn assuming we have infinite data?
  - Learning curve
    - Rate of convergence (in amount of training data) toward asymptotic (infinite data) hypothesis
Gaussian Naïve Bayes vs. Logistic Regression

- Consider boolean $Y$, continuous $X_i$'s
- Number of parameters to estimate
  - GNB
  - GNB2
  - LR
Gaussian Naïve Bayes vs. Logistic Regression

- Consider boolean Y, continuous $X_i$'s
- Number of parameters to estimate
  - GNB: $4n+1$
  - GNB2: $3n+1$
  - LR: $n+1$

- Estimation method
  - NB parameter estimates are uncoupled
  - LR parameter estimates are coupled
Case study

- Assume \( Y = \text{PlayBasketball} \) (boolean) \( X_1 = \text{Height} \) \( X_2 = \text{Age} \)
  \[ Y_{\text{New}} \leftarrow \arg \max_{y_k} P(Y | y_k) \prod_{y_k} P(X_i^{\text{New}} | Y = y_k) \]; assume \( P(Y=1) = 0.5 \)
Gaussian Naïve Bayes vs. Logistic Regression

- Recall the two assumptions while deriving the form of LR from GNB
  - 1. $X_i$ are conditionally independent of $X_k$ given $Y$
  - 2. $P(X_i | Y = y_k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$; NOT $\mathcal{N}(\mu_{ik}, \sigma_{ik})$

- Consider three learning methods:
  - GNB (assumption 1 only)
  - GNB2 (assumption 1 and 2)
  - LR

- Which method works better if we have infinite training data and
  - Both (1) and (2) are satisfied
  - Neither (1) nor (2) is satisfied
  - (1) is satisfied but not (2)
Gaussian Naïve Bayes vs. Logistic Regression

○ Recall the two assumptions while deriving the form of LR from GNB
  ○ 1. $X_i$ are conditionally independent of $X_k$ given $Y$
  ○ 2. $P(X_i \mid Y = y_k) \sim \mathcal{N} (\mu_{ik}, \sigma_i)$; NOT $\mathcal{N} (\mu_{ik}, \sigma_{ik})$

○ Consider three learning methods:
  ○ GNB (assumption 1 only)
  ○ GNB2 (assumption 1 and 2)
  ○ LR

○ Which method works better if we have infinite training data and

  ○ Both (1) and (2) are satisfied $\quad$ LR = GNB2 = GNB
  ○ Neither (1) nor (2) is satisfied $\quad$ LR > GNB2, GNB > GNB2
  ○ (1) is satisfied but not (2) $\quad$ GNB > LR, LR > GNB2
Gaussian Naïve Bayes vs. Logistic Regression

- What if we have finite training data?

- GNB and LR converge at different rates to asymptotic ($\infty$ data) error

- Let $\varepsilon_{A,n}$ refer to expected error of learning algorithm A after n training examples

- Let $d$ be the number of features $<X_1, X_2, ..., X_d>$

  - $\varepsilon_{LR,n} = \varepsilon_{LR,\infty} + O\left(\sqrt{\frac{d}{n}}\right)$

  - $\varepsilon_{GNB,n} = \varepsilon_{GNB,\infty} + O\left(\sqrt{\frac{\log d}{n}}\right)$

- So GNB requires $d = O(\log d)$ to converge, but LR requires $d = O(d)$
Naïve Bayes vs. Logistic Regression

- The bottom line
  - GNB2 and LR both use linear decision surface, GNB need not
  - Given infinite training data, LR is better than GNB2 because the training is free from assumptions (although our derivation of the form of $P(Y|X)$ did)
  - But GNB2 converges more quickly to perhaps less-accurate asymptotic error
  - And GNB is more biased (assumption 1) and less (assumption 2) than LR, so neither might beat each other.