CS 4824/ECE 4424: 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.

Let's learn classifiers by learning P(Y|X)

- Suppose we want to learn the function f: $<G, H> \rightarrow W$
- Or $P(W \mid G, H)$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

- Suppose $X = \langle X_1, X_2, \ldots, X_n \rangle$, where X_i and Y are boolean RV

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

- To estimate $P(Y \mid X_1, X_2, \ldots, X_n)$, how many parameters do we need?

- How can we design a learning algorithm that is practical?

- Can Bayes Rule help?
  - $$P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)}$$

Can we reduce parameters using Bayes Rule?

- Suppose $X = \langle X_1, X_2, \ldots, X_n \rangle$, where X_i and Y are boolean RV
- Bayes Rule:
  - $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$
- How many parameters needed to estimate $P(X_1, X_2, \ldots, X_n | Y)$?
- How many parameters needed to estimate $P(Y) = 1$?

Naïve Bayes

- Naïve Bayes assumes
  - $$P(X_1, \ldots, X_n \mid Y) = \prod_i P(X_i \mid Y)$$
- That is, X_i and X_j are conditionally independent given Y $\forall i \neq j$

Conditional independence

- X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y given the value of Z
  - $(\forall i, j, k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$
  - Or equivalently, P(X|Y Z) = P(X|Z)
- P(Thunder|Rain, Lightning) = P(Thunder|Lightning)
  - That is, Thunder and Rain are conditionally independent

Naïve Bayes assumes conditional independence

- Under the conditional independence assumption, then
  - $P(X_1, X_2 \mid Y) =$

Naïve Bayes assumes conditional independence

- In General,
  - $$P(X_1, \ldots, X_n \mid Y) = \prod_i P(X_i \mid Y)$$
- How many parameters to describe $P(X_1, \ldots, X_n \mid Y)$? $P(Y)$?
  - Without conditional independence:
  - With conditional independence:

Naïve Bayes summary

- 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 $X^{New} = \langle X_1, X_2, \ldots, X_n \rangle$?

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 $\theta_{ijk} = P(X = x_{ij} \mid Y = y_k)$

- Classify X^{New}
  - $Y^{New} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{New} \mid Y = y_k)$
  - $Y^{New} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$

How to estimate parameters: discrete-valued Y, X_i

- Maximum likelihood estimates (MLE's)
  - $$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$
  - $$\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} | Y = y_k) = \frac{\#D\{X_j = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

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 $X_i=X_k$
    - $P(Y=1|X) = P(Y=1)\, P(X_1|Y=1)\, P(X_2|Y=1) \ldots P(X_i|Y=1) \ldots P(X_k|Y=1)$

Naïve Bayes issue #2

- If unlucky, the MLE estimate for $P(X_i|Y)$ 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?
  - $$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$
  - $$\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} | Y = y_k) = \frac{\#D\{X_j = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Using MAP estimation: discrete-valued Y, X_i

- Maximum a posteriori estimate (MAP)
  - (Beta, Dirichlet prior)
  - $$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$
  - $$\hat{\theta}_{ijk} = \hat{P}(X = x_{ij} | Y = y_k) = \frac{\#D\{X_j = x_{ij} \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

Questions to think about…

What's the decision rule of Naïve Bayes?

What if we have continuous X_i?