UMBC

AN HONORS UNIVERSITY IN MARYLAND

Computational Methods in IS Research
Spring 2016

Bayesian Learning

Nirmalya Roy

Department of Information Systems

University of Maryland Baltimore County

Bayesian Learning

- Combines prior knowledge with evidence to make predictions
- Optimal (albeit impractical) classifier
- Naïve Bayes classifier (practical)
  - Assumes independence among features
- Association rule mining

Bayes Rule

Thomas Bayes

$$P(C_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_i)P(C_i)}{p(\boldsymbol{x})}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

- C_i is the class, $1 \le i \le K$
- $\boldsymbol{x}$ is the feature vector of an instance
- $P(C_i \mid \boldsymbol{x})$ = probability that instance $\boldsymbol{x}$ belongs to class C_i (*posterior*)
- $p(\boldsymbol{x} \mid C_i)$ = probability that an instance drawn from class C_i would be $\boldsymbol{x}$ (*likelihood*)
- $P(C_i)$ = probability of class C_i (*prior*)
- $p(\boldsymbol{x})$ = probability of instance $\boldsymbol{x}$ (*evidence*)

Intuition behind different Probabilities

- *Prior probability:*
  - Knowledge we have as to the value of C before looking at observables $\boldsymbol{x}$
- *Likelihood probability:*
  - Conditional probability that an event belonging to C has the associated observation value $\boldsymbol{x}$
  - Data tells us regarding the class
- *Evidence:*
  - Marginal probability that an observation $\boldsymbol{x}$ is seen

Bayes Classifier

$$P(C_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_i)P(C_i)}{p(\boldsymbol{x})}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

- Classify instance $\boldsymbol{x}$ as class C_i such that

$$i = \arg\max_{1 \le k \le K} P(C_k \mid \boldsymbol{x})$$

posterior

- Since only interested in maximum, can ignore denominator $p(\boldsymbol{x})$

$$i = \arg\max_{1 \le k \le K} p(\boldsymbol{x} \mid C_k)P(C_k)$$

- If prior probability distribution of classes is uniform, then can ignore $P(C_i)$

$$i = \arg\max_{1 \le k \le K} p(\boldsymbol{x} \mid C_k)$$

Bayes Classifier

- Practical issue
  - $p(\mathbf{x}|\, C_i)$ is a joint probability distribution
  - Need to know the probability of every possible instance given every possible class
  - Even for D boolean features and K classes, that's $K*2^D$ probabilities
- Solution
  - Assume features are independent of each other

$$p(x_1, x_2, \ldots, x_D \mid C_i) = \prod_{j=1}^{D} p(x_j \mid C_i)$$

Naïve Bayes Classifier

- Given training set X
- Estimate probabilities from X

$$P(C_i) = \frac{|\{(\boldsymbol{x}, r) \in X \mid r = C_i\}|}{|X|}$$

$$p(x_j = v \mid C_i) = \frac{|\{(\boldsymbol{x}, r) \in X \mid x_j = v \, and \, r = C_i\}|}{|\{(\boldsymbol{x}, r) \in X \mid r = C_i\}|}$$

- Classify new instance $\boldsymbol{x}$ as class C_i such that

$$i = \arg\max_{1 \le k \le K} P(C_k) * \prod_{j=1}^{D} p(x_j \mid C_k)$$

Naïve Bayes Classifier

- Another practical issue
  - What if x_j is a continuous feature?
- Solution #1
  - Assume some parameterized distribution for x_j
    - E.g., normal
  - Learn parameters of distribution from data
    - E.g., mean and variance of x_j values
- Solution #2
  - Discretize feature
  - E.g., price $\in \Re$ to price $\in$ {low, medium, high}

Naïve Bayes Classifier

- Yet another practical issue
  - What if no examples in class C_i have $x_j = v$?

$$p(x_j = v \mid C_i) = 0$$

$$P(C_i) * \prod_{j=1}^{D} p(x_j \mid C_i) = 0$$

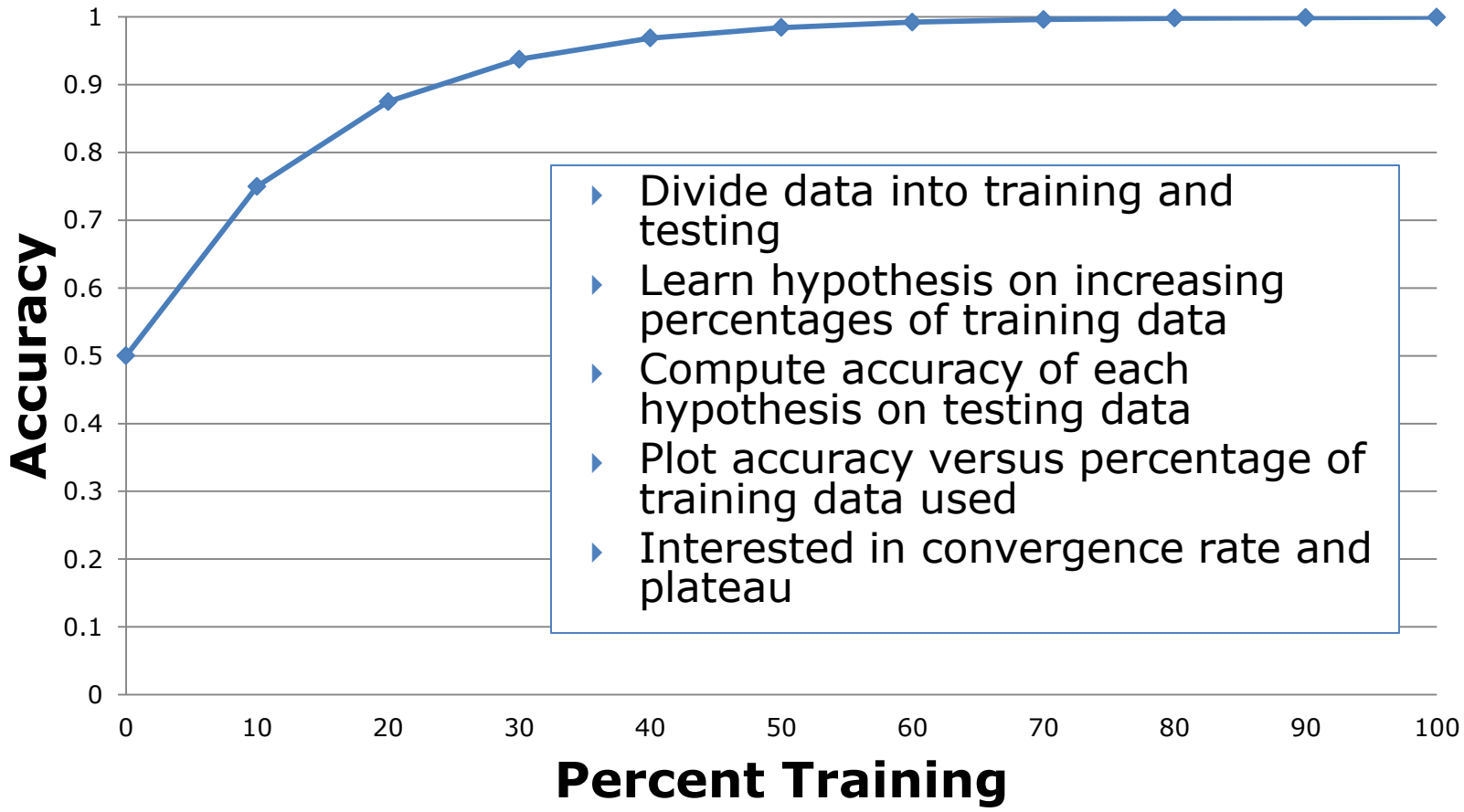
- Solution

$$p(x_j = v \mid C_i) = \frac{|\{(\mathbf{x}, r) \in X \mid x_j = v \, and \, r = C_i\}| + 1}{|\{(\mathbf{x}, r) \in X \mid r = C_i\}| + |domain(x_j)|}$$

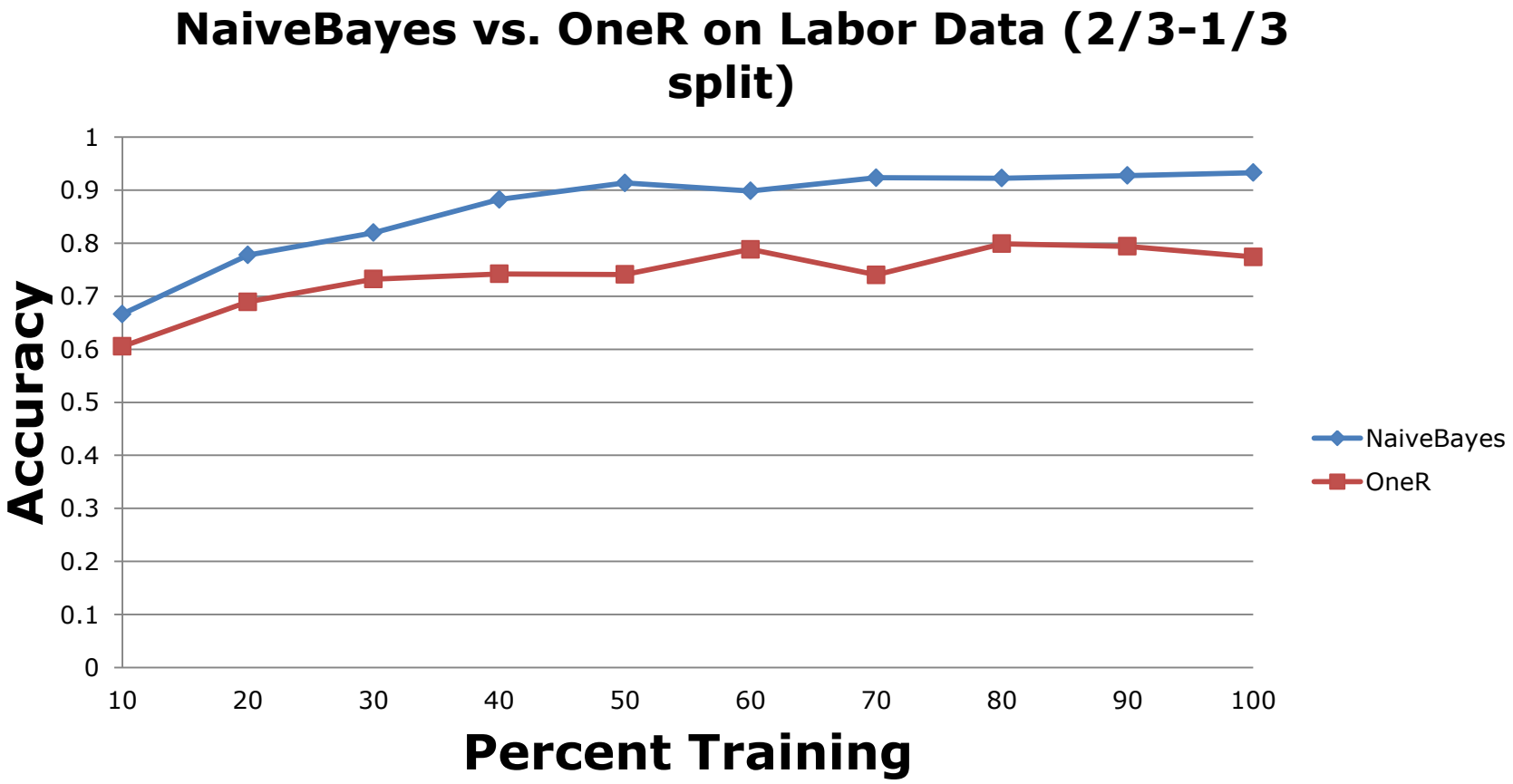
Naïve Bayes Classifier

- Independence assumption rarely true
  - E.g., is “price” independent from “engine power”?
- Naïve Bayes classifier still does surprisingly well
- Simple, effective baseline for other learners

Sidebar: The Learning Curve



Learning Curve



Association Rules

- Association rule: $X \rightarrow Y$
  - *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y*

- A rule implies association, not necessarily causation

Association measures

- Support ($X \rightarrow Y$):

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence ($X \rightarrow Y$):

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

- Lift ($X \rightarrow Y$):

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

$$= \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

Significance of Association measures

- *Confidence*
  - Conditional probability
  - Value should be close to 1
  - Strength of the rule, rule holds enough confidence
- *Support*
  - Statistical significance of the rule
  - Strong confidence value but # of such customers is small, rule is worthless
- Minimum support and confidence are set by the user/entity
  - Rules with higher support and confidence are searched for in the database

Association Rules

- In general, X and Y can be a sets of items
  - Basket analysis
  - E.g., customers buying hot dogs and buns are more likely to buy mustard and catsup
- Association rule mining
  - Given database of customer purchases
  - Find all association rules with high support and confidence
  - Apriori algorithm [Agrawal et al., 1996]

Apriori Algorithm (Agrawal et al., 1996)

- For (X,Y,Z), a 3-item set, to be frequent (have enough support), (X,Y), (X,Z), and (Y,Z) should be frequent
- If (X,Y) is not frequent, none of its supersets can be frequent
- Once we find the frequent *k*-item sets, we convert them to rules: X, Y $\rightarrow$ Z, ...

  and X $\rightarrow$ Y, Z, ...

Apriori Algorithm

- Find all itemsets with enough support
  - If itemset of size k does not have enough support, then no superset of this itemset will have enough support
- For each itemset, find all association rules X→Y with enough confidence
  - Rules of the form {A,B} → {C,D} can only be confident if both {A,B,C} → {D} and {A,B,D} → {C} are confident
- WEKA: “Associate”

Summary: Bayesian Learning

- Optimal learning framework
- Incorporates background knowledge
- Practical algorithms (Naïve Bayes)
- Association rule mining