MLE's, Bayesian Classifiers and Naïve Bayes

Required reading:

• Mitchell draft chapter (on class website)

Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 28, 2008

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Example: Bernoulli model

- Data:
  - We observed N ***iid*** coin tossing: D={1, 0, 1, …, 0}

- Representation:

Binary r.v: $x_n = \{0,1\}$

$x = 1 \Rightarrow \theta$

- Model:

$$P(x) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$$

- How to write the likelihood of a single observation x_i?

$$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

$\arg\max_{\theta} P(data|\theta)$

- The likelihood of dataset $D=\{x_1, \ldots, x_N\}$:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N}\left(\theta^{x_i}(1-\theta)^{1-x_i}\right) = \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}}(1-\theta)^{\#\text{tails}}$$

Estimating MLE for Bernoulli model

$$P(x_1, \dots x_n | \theta) = \prod_i \theta^{x_i} (1-\theta)^{(1-x_i)} \qquad \frac{\partial \log(x)}{\partial x} = \frac{1}{x}$$

$$\log P(\quad | \theta) = \sum_i \log(\quad) = \sum_i \left[x_i \log \theta + (1-x_i) \log(1-\theta) \right]$$

$$\frac{\partial}{\partial \theta} \log P(x_1 \cdots x_n | \theta) = \sum_i x_i \left(\frac{1}{\theta}\right) + \sum_i (1-x_i) \frac{1}{1-\theta} (-1) = 0$$

$$\frac{1}{\theta} \underbrace{\sum_i x_i}_{\#\text{heads}} = \frac{1}{1-\theta} \underbrace{\sum_i (1-x_i)}_{\#\text{tails}}$$

$$1-\theta \sum_i x_i = \theta \sum_i (1-x_i)$$

$$\sum_i x_i - \theta \sum_i x_i = \theta \sum_i 1 - \theta \sum_i x_i$$

$$\frac{\sum_i x_i}{\sum_i 1} = \theta$$

Naïve Bayes and Logistic Regression

- Design learning algorithms based on probabilistic model
  - Learn f: X → Y, or better yet P(Y|X)
- Two of the most widely used algorithms
- Interesting relationship between these two:
  - Generative vs Discriminative classifiers

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Random Variable

ith possible value of Y

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Common abbreviation:

$$(\forall i, j) \;\; P(y_i | x_j) = \frac{P(x_j | y_i) P(y_i)}{P(x_j)}$$

Bayes Classifier

Training data:

X (Sky … Forecst) $\quad Y$ (EnjoySpt)

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$P(Y_i=1 \mid x_i)$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$P(x) = \sum_i P(x|y_i)P(y_i)$

Learning = estimating P(X|Y), P(Y)

Classification = using Bayes rule to calculate P(Y | X^{new})

Bayes Classifier

Training data:

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

(X = Sky … Forecst; Y = EnjoySpt)

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

How shall we represent P(X|Y), P(Y)?

How many parameters must we estimate?

Bayes Classifier

Training data:

X (Sky … Forecst) Y (EnjoySpt)

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

How shall we r… P(X|Y), P(Y)?

How man… …eters must we estimate?

Full joint $P(X_1 \ldots X_n \mid Y)$ usually impractical!

Naïve Bayes

Naïve Bayes assumes

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

Conditional Independence

Definition: X is <u>conditionally independent</u> of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y,Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y

$P(X_1 \cdots X_{21} | Y)$ — 2×2^{21} if no cond indep — 42 if cond indep

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

= by cond indep of X_1, X_2 given Y

in general: $P(X_1...X_n|Y) = \prod_i P(X_i|Y)$

How many parameters needed to describe P(X|Y)? P(Y)?

- Without conditional indep assumption?
- With conditional indep assumption?

$P(X_i=1 | Y=1)$
$P(X_i=0 | Y=1) = 1 -$
$P(X_i=0 | Y=0)$
$P(X_i=1 | Y=0)$

How many parameters to estimate?

P(X1, ... Xn | Y), all variables boolean

Without conditional independence assumption:

With conditional independence assumption:

Naïve Bayes in a Nutshell

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:

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

So, classification rule for $X^{new} = \langle X_1, \ldots, X_n \rangle$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij}|Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new}|Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} \; P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} \; P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} \; = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

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

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in set D for which $Y=y_k$