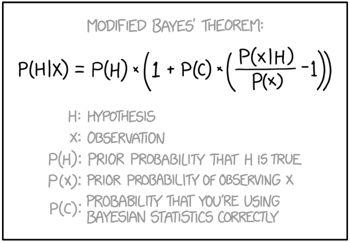


Image. xkcd comics - # 2059

Data Mining: Classification: Part 3

Naive Bayes

CS 4821 - CS 5831 - s24

Some slides adapted from P. Smyth; A. Moore, D. Klein Han, Kamber, Pei; Tan, Steinbach, Kumar; E. Keogh; Z. Bar-Joseph; L. Kaebling; R. Tibshirani; T. Taylor; and L. Hannah

Review of Probability

Probability Review

Probability on a set is defined by three basic elements:

- **Sample Space** Ω: the set of all outcomes of a random experiment. An outcome $\omega \in \Omega$ completely describes the state of the world
- **Set of Events** $\mathcal{F}$: a set with elements $A \in \mathcal{F}$ are subsets of Ω
- **Probability measure**: A function $P : \mathcal{F} \to \mathbb{R}$ satisfying three basic axioms

Axioms of Probability

1. $P(A) \geq 0$, for all $A \in \mathcal{F}$
   $P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   if $A_1, A_2, \ldots$ are disjoint events, then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

We will use random variables X a function $X : \Omega \to \mathbb{R}$, and look at probability of a set associated with a random variable X taking on a value x

$$P(X = x)$$

Joint Distributions

- A joint distribution over a set of random variables $X_1, X_2, \ldots, X_p$ specifies a real number for each assignment

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

- Joint distributions must obey $P(x_1, x_2, \ldots, x_p) \geq 0$ and $\sum_{x_1, x_2, \ldots, x_p} P(x_1, x_2, \ldots, x_p) = 1$
- Distribution can be represented with a matrix or table

	alarm	¬ alarm
burglary	0.09	0.01
¬ burglary	0.10	0.80

Alarm	Burglary	Prob.
a	b	0.09
$\neg a$	b	0.01
a	$\neg b$	0.10
$\neg a$	$\neg b$	0.80

Relationships with Probability

- Conditional Probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A, B)}{P(A)}$$

- Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Independence

- Two variables are independent if:

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

$$\forall x_1, x_2 \; P(x_1, x_2) = P(x_1)P(x_2)$$

- Example. X_1 is independent of X_2, then

$$P(X_1 \mid X_2) = P(X_1)$$

$$\begin{aligned} P(\neg X_1 \mid X_2) &= P(\neg X_1) & P(\neg X_2, X_1) &= P(\neg X_2)P(X_1) \\ P(X_2 \mid X_1) &= P(X_2) & P(X_2, \neg X_1) &= P(X_2)P(\neg X_1) \\ P(X_2, X_1) &= P(X_2)P(X_1) & P(\neg X_2, \neg X_1) &= P(\neg X_2)P(\neg X_1) \end{aligned}$$

Independence is denoted: $X_1 \perp X_2$

Conditional Probability

- Consider conditional independence of X_1 and X_2 given Y denoted: $X_1 \perp X_2 \mid Y$

$$\forall x_1, x_2, y \; P(x_1 \mid x_2, y) = P(x_1 \mid y)$$

$$P(X_1 \mid X_2, Y) = P(X_1 \mid Y)$$

$$P(X_1, X_2 \mid Y) = P(X_1 \mid Y) P(X_2 \mid Y)$$

Example of Bayes Theorem

- Given:
  - A doctor knows meningitis causes stiff necks 50% of the time
  - Prior probability of any patient having meningitis is $\frac{1}{50000}$
  - Prior probability of any patient having a stiff neck is $\frac{1}{20}$
- If a patient has a stiff neck, what is the probability they have meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 * \frac{1}{50000}}{\frac{1}{20}} = 0.0002$$

Bayes Theorem

- Given training data $\mathbf{X}$, calculate the *posteriori* probability of a hypothesis H, $P(H \mid \mathbf{X})$ via Bayes theorem

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})}$$

- Informally, this is

$$\text{posteriori} = \text{likelihood} \times \text{prior}/\text{evidence}$$

- In Classification with Naïve Bayes, we use this relationship to predict a sample belongs to a class $y_i \in Y$.
  - if $P(y_i \mid \mathbf{X})$ is the largest among all the $P(y_k \mid \mathbf{X})$ for all k, then the Naïve Bayes classifier with predict $\hat{y}_i$

$$\hat{y} = \arg\max\nolimits_k P(y_k \mid \mathbf{X}) = \arg\max_k P(y_k \mid X_1, X_2, \ldots, X_p)$$

Further Review of Probability

Available in slides: 03.classify.extra.probability.review.pdf

Bayesian Classifier

Simple Example

- Two binary variables (X_1, X_2) and class label, Y, with simple application of Bayes rule:

$$P(y \mid X_1, X_2) = \frac{P(X_1, X_2 \mid y)P(y)}{P(X_1, X_2)}$$

Bayes Estimate (no assumptions)

$$P(y \mid X_1, X_2) = \frac{P(X_1 \mid y)P(X_2 \mid y)P(y)}{P(X_1, X_2)}$$

Naïve Bayes (assume conditional independence)

General Naïve Bayes

$$P(y_k \mid X_1, X_2, \ldots, X_p) = \frac{P(X_1, X_2, \ldots, X_p \mid y_k)P(y_k)}{P(X_1, X_2, \ldots, X_p)}$$

Assume conditional independence of all variables given the class

$$X_1 \perp X_2 \,|\, Y, \quad X_1 \perp X_3 \,|\, Y, \quad \ldots, \quad X_{p-1} \perp X_p \,|\, Y$$

$$P(y_k \mid X_1, X_2, \ldots, X_p) = \frac{\prod_j P(X_j \mid y_k)P(y_k)}{P(X_1, X_2, \ldots, X_p)}$$

$$P(y_k \mid X_1, X_2, \ldots, X_p) \propto \prod_j P(X_j \mid y_k)P(y_k)$$

General Naïve Bayes

The Naïve Bayes classifier returns a class label of:

$$\hat{y}^{NB} = \arg\max_{k} \; P(Y = y_k) \prod_{i=1}^{p} P(X_i = x_{test,i} \mid Y = y_k)$$

How to estimate the probabilities?

Let's look at an example.

Naïve Bayes Example 1

Naïve Bayes Example 1

Want to predict whether you Play a tennis match (Y / N) given four factors:

- Outlook - { Sunny, Overcast, Rain }
- Temperature - { Hot, Mild, Cold }
- Humidity - { High, Low }
- Windy - { True, False }

First item in Naïve Bayes, estimate probabilities

- Prior probabilities: $P(Play = y)$, $P(Play = n)$
- Conditional probabilities: $P(Outlook \,|\, Play)$, $P(Temp \,|\, Play)$, $\ldots$

Naïve Bayes Example 1

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	Low	True	Yes
Sunny	Mild	Low	True	Yes
Rainy	Mild	High	False	Yes
Overcast	Cool	High	False	Yes
Rainy	Cool	High	True	No
Overcast	Cool	Low	True	No
Sunny	Cool	High	True	Yes
Rainy	Mild	High	True	No
Rainy	Cool	Low	False	Yes
Overcast	Hot	Low	True	Yes
Sunny	Cool	Low	False	Yes
Overcast	Mild	High	False	Yes

Maximum likelihood estimates

$$\hat{P}(y_k) = \frac{\#(Y = y_k)}{n} = \frac{n_k}{n} \qquad \hat{P}(x_i \mid y_k) = \frac{\#(X_i = j, Y = y_k)}{\#(Y = y_k)} = \frac{n_{ijk}}{n_k}$$

where n_{ijk} - num. of records with $Y = y_k, X_i = j$ and n_k - num. of records with $Y = y_k$

Example 1: Naïve Bayes Estimate Probabilities

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	Low	True	Yes
Sunny	Mild	Low	True	Yes
Rainy	Mild	High	False	Yes
Overcast	Cool	High	False	Yes
Rainy	Cool	High	True	No
Overcast	Cool	Low	True	No
Sunny	Cool	High	True	Yes
Rainy	Mild	High	True	No
Rainy	Cool	Low	False	Yes
Overcast	Hot	Low	True	Yes
Sunny	Cool	Low	False	Yes
Overcast	Mild	High	False	Yes

$P(Play)$
$P(Play = N) = 4/13$
$P(Play = Y) = 9/13$

$P(Outlook = \{R, O, S\} \mid Play)$	
$P(R \mid N) = 2/4$	$P(R \mid Y) = 2/9$
$P(O \mid N) = 1/4$	$P(O \mid Y) = 3/9$
$P(S \mid N) = 1/4$	$P(S \mid Y) = 4/9$

$P(Temp = \{H, M, C\} \mid Play)$	
$P(H \mid N) = 1/4$	$P(H \mid Y) = 2/9$
$P(M \mid N) = 1/4$	$P(M \mid Y) = 3/9$
$P(C \mid N) = 2/4$	$P(C \mid Y) = 4/9$

$P(Humidity = \{H, L\} \mid Play)$	
$P(H \mid N) = 3/4$	$P(H \mid Y) = 4/9$
$P(L \mid N) = 1/4$	$P(L \mid Y) = 5/9$

$P(Windy = \{F, T\} \mid Play)$	
$P(F \mid N) = 1/4$	$P(F \mid Y) = 5/9$
$P(T \mid N) = 3/4$	$P(T \mid Y) = 4/9$

Example 1: Naïve Bayes Prediction

Given the estimated probabilities, determine which class (No / Yes) does a new data sample maximize the probabilities.

For the test data sample (Rain, Hot, High, False), need to calculate two values:

- $P(\text{Play}{=}\text{N} \mid \text{Rain, Hot, High, False})$
  $\propto P(\text{Rain} \mid N)P(\text{Hot} \mid N)P(\text{High} \mid N)P(\text{False} \mid N)P(N)$

- $P(\text{Play}{=}\text{Y} \mid \text{Rain, Hot, High, False})$
  $\propto P(\text{Rain} \mid Y)P(\text{Hot} \mid Y)P(\text{High} \mid Y)P(\text{False} \mid Y)P(Y)$

Example 1: Naïve Bayes

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	Low	True	Yes
Sunny	Mild	Low	True	Yes
Rainy	Mild	High	False	Yes
Overcast	Cool	High	False	Yes
Rainy	Cool	High	True	No
Overcast	Cool	Low	True	No
Sunny	Cool	High	True	Yes
Rainy	Mild	High	True	No
Rainy	Cool	Low	False	Yes
Overcast	Hot	Low	True	Yes
Sunny	Cool	Low	False	Yes
Overcast	Mild	High	False	Yes

$P(Play)$
$P(Play = N) = 4/13$
$P(Play = Y) = 9/13$

$P(Outlook = \{R, O, S\} \mid Play)$	
$P(R \mid N) = 2/4$	$P(R \mid Y) = 2/9$
$P(O \mid N) = 1/4$	$P(O \mid Y) = 3/9$
$P(S \mid N) = 1/4$	$P(S \mid Y) = 4/9$

$P(Temp = \{H, M, C\} \mid Play)$	
$P(H \mid N) = 1/4$	$P(H \mid Y) = 2/9$
$P(M \mid N) = 1/4$	$P(M \mid Y) = 3/9$
$P(C \mid N) = 2/4$	$P(C \mid Y) = 4/9$

$P(Humidity = \{H, L\} \mid Play)$	
$P(H \mid N) = 3/4$	$P(H \mid Y) = 4/9$
$P(L \mid N) = 1/4$	$P(L \mid Y) = 5/9$

$P(Windy = \{F, T\} \mid Play)$	
$P(F \mid N) = 1/4$	$P(F \mid Y) = 5/9$
$P(T \mid N) = 3/4$	$P(T \mid Y) = 4/9$

$$P(Play = N \mid \text{Rain, Hot, High, False})$$
$$\propto P(R \mid N) P(H \mid N) P(H \mid N) P(F \mid N) P(N)$$
$$= 2/4 * 1/4 * 3/4 * 1/4 * 4/13 = 0.007212$$

Example 1: Naïve Bayes

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	Low	True	Yes
Sunny	Mild	Low	True	Yes
Rainy	Mild	High	False	Yes
Overcast	Cool	High	False	Yes
Rainy	Cool	High	True	No
Overcast	Cool	Low	True	No
Sunny	Cool	High	True	Yes
Rainy	Mild	High	True	No
Rainy	Cool	Low	False	Yes
Overcast	Hot	Low	True	Yes
Sunny	Cool	Low	False	Yes
Overcast	Mild	High	False	Yes

$P(Play)$
$P(Play = N) = 4/13$
$P(Play = Y) = 9/13$

$P(Outlook = \{R, O, S\} \mid Play)$	
$P(R \mid N) = 2/4$	$P(R \mid Y) = 2/9$
$P(O \mid N) = 1/4$	$P(O \mid Y) = 3/9$
$P(S \mid N) = 1/4$	$P(S \mid Y) = 4/9$

$P(Temp = \{H, M, C\} \mid Play)$	
$P(H \mid N) = 1/4$	$P(H \mid Y) = 2/9$
$P(M \mid N) = 1/4$	$P(M \mid Y) = 3/9$
$P(C \mid N) = 2/4$	$P(C \mid Y) = 4/9$

$P(Humidity = \{H, L\} \mid Play)$	
$P(H \mid N) = 3/4$	$P(H \mid Y) = 4/9$
$P(L \mid N) = 1/4$	$P(L \mid Y) = 5/9$

$P(Windy = \{F, T\} \mid Play)$	
$P(F \mid N) = 1/4$	$P(F \mid Y) = 5/9$
$P(T \mid N) = 3/4$	$P(T \mid Y) = 4/9$

$$P(Play = Y \mid \text{Rain, Hot, High, False})$$
$$\propto P(R \mid Y) P(H \mid Y) P(H \mid Y) P(F \mid Y) P(Y)$$
$$= 2/9 * 2/9 * 4/9 * 5/9 * 9/13 = 0.008441$$

Example 1: Naïve Bayes Prediction

For the test data sample (Rain, Hot, High, False), need to calculate two values:

- $P(\text{Play=N} \mid \text{Rain, Hot, High, False})$
  $\propto P(\text{Rain} \mid N)P(\text{Hot} \mid N)P(\text{High} \mid N)P(\text{False} \mid N)P(N) = 0.007212$

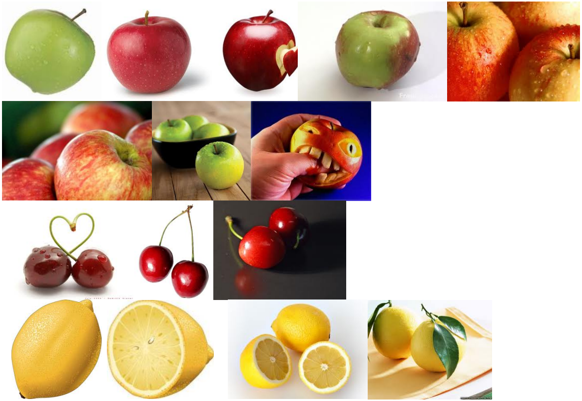
- $P(\text{Play=Y} \mid \text{Rain, Hot, High, False})$
  $\propto P(\text{Rain} \mid Y)P(\text{Hot} \mid Y)P(\text{High} \mid Y)P(\text{False} \mid Y)P(Y) = 0.008441$

Because
$P(Y) \mid \text{Rain, Hot, High, False}) > P(N \mid \text{Rain, Hot, High, False})$
label sample as <u>Play = Yes</u>

Naïve Bayes Example 2

Example 2: Naïve Bayes

Use Naïve Bayes to predict fruit given a few attributes:
Color, Shape, Size[1]



{Example from L. Hannah}

Example 2: Naïve Bayes

Need to estimate the following:

- Class probabilities: $P(apple)$, $P(cherry)$, $P(lemon)$
- Feature conditional probabilities given the class: $P(green \mid apple)$, $P(red \mid apple)$, $\ldots$

Test on:

Example 2: Training Data

Color	Shape	Size	Fruit
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Example 2: Estimate Probabilities

Class probabilities:

- $P(apple) =$
- $P(cherry) =$
- $P(lemon) =$

Color	Shape	Size	Fruit
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Example 2: Estimate Probabilities

Conditional probabilities

- $P(red \mid apple) =$
- $P(green \mid apple) =$
- $P(yellow \mid apple) =$

Color	**Shape**	**Size**	**Fruit**
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Example 2: Estimate Probabilities

Conditional probabilities

- $P(red \mid apple) =$

- $P(green \mid apple) =$

- $P(yellow \mid apple) =$

Color	Shape	Size	Fruit
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Issue! $P(yellow \mid apple) = 0$, but test sample is ...

Smoothing

Idea: change estimate of probabilities, so not equal to 0

- Maximum likelihood estimates

$$\hat{P}(y_k) = \frac{\#(Y = y_k)}{n} = \frac{n_k}{n} \qquad \hat{P}(x_i \mid y_k) = \frac{\#(X_i = j, Y = y_k)}{\#(Y = y_k)} = \frac{n_{ijk}}{n_k}$$

  where n_{ijk} - num. of records with $Y = y_k, X_i = j$ and n_k - num. of records with $Y = y_k$

- Smoothing - Laplace
  Adjust estimates of probability:

$$\hat{P}(x_i \mid y_k) = \frac{n_{ijk} + \alpha}{n_k + \beta}$$

  where α and β are parameters.
  - Let's consider $\alpha = 1$, β is the num. of values of X_i

Example 2: Estimate Probabilities

Let $\alpha = 1, \beta = \#$ colors. Compute the conditional probabilities

- $P(red \mid apple)$, $P(green \mid apple)$, $P(yellow \mid apple)$
- $P(red \mid cherry)$, $P(green \mid cherry)$, $P(yellow \mid cherry)$
- $P(red \mid lemon)$, $P(green \mid lemon)$, $P(yellow \mid lemon)$

Color	**Shape**	**Size**	**Fruit**
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Example 2: Estimate Probabilities

Let $\alpha = 1, \beta = \#$ shapes. Compute the conditional probabilities

- $P(round \mid apple)$, $P(oval \mid apple)$
- $P(round \mid cherry)$, $P(oval \mid cherry)$
- $P(round \mid lemon)$, $P(oval \mid lemon)$

Color	**Shape**	**Size**	**Fruit**
Green	Round	2.1	Apple
Red	Round	1.9	Apple
Red	Round	2.0	Apple
Green	Round	1.8	Apple
Red	Round	1.9	Apple
Red	Round	2.1	Apple
Green	Round	1.6	Apple
Red	Round	1.7	Apple
Red	Round	1.1	Cherry
Red	Round	1.0	Cherry
Red	Round	1.2	Cherry
Yellow	Oval	2.8	Lemon
Yellow	Oval	2.6	Lemon
Yellow	Oval	2.5	Lemon
Yellow	Round	2.7	Lemon

Example 2: Estimate Probabilities

For the conditional size probabilities, have a continuous variable:

- bin sizes to make discrete data
  $\{size < 2\}, \{2 \leq size > 2.5\}, \{size \geq 2.5\}$
- model probabilities as Gaussian with mean $\hat{\mu}$ and $\hat{\sigma}^2$

$$P(X_i \mid y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{|x_i - \mu_{ik}|^2}{2\sigma_{ik}^2}}$$

Example 2: Training Data

Color	**Shape**	**Size**	**Fruit**
Green	Round	Medium	Apple
Red	Round	Small	Apple
Red	Round	Medium	Apple
Green	Round	Small	Apple
Red	Round	Small	Apple
Red	Round	Medium	Apple
Green	Round	Small	Apple
Red	Round	Small	Apple
Red	Round	Small	Cherry
Red	Round	Small	Cherry
Red	Round	Small	Cherry
Yellow	Oval	Large	Lemon
Yellow	Oval	Large	Lemon
Yellow	Oval	Large	Lemon
Yellow	Round	Large	Lemon

Example 2: Predict type of fruit

Which fruit is this?

Color = yellow, shape = round, size = 1.8

Example 2: Predict type of fruit

Compute:

- $P(apple \mid yellow, round, size < 2) \propto$
- $P(cherry \mid yellow, round, size < 2) \propto$
- $P(lemon \mid yellow, round, size < 2) \propto$

Maximum value is the predicted class

Naïve Bayes Summary

Naïve Bayes in Practice - Estimate Probabilities

Issue:

As seen in example 2, the estimates of the probabilities can be driven to 0 or 1 with small training data

$$\hat{P}(x_i \mid y_k) = \frac{\#(X_i = j, Y = y_k)}{\#(Y = y_k)} = \frac{n_{ijk}}{n_k}$$

where n_{ijk} - num. of records with $Y = y_k, X_i = j$ and n_k - num. of records with $Y = y_k$

Naïve Bayes in Practice - Estimate Probabilities

Solution:

Use smoothing, e.g., Laplace

Different smoothers:

$$P(y_k) = \frac{n_k + 1}{n + |Y|} \quad \text{or} \quad \frac{n_k + m}{n + m|Y|}$$

$$\hat{P}(x_i \mid y_k) = \frac{n_{ijk} + 1}{n_k + |X_i|} \quad \text{or} \quad \frac{n_{ijk} + m}{n_k + m|X_i|}$$

where m is a hyper-parameter.

Naïve Bayes in Practice - Underflow Error

Problem:

Multiplying a bunch of small numbers, can get underflow errors

$$\prod_j P(x_j \mid y_k)P(y_k) = P(x_1 \mid y_k)P(x_2 \mid y_k) \cdots P(x_p \mid y_k)P(y_k)$$

Naïve Bayes in Practice - Underflow Error

Solution:

Calculate $\log$ of the probabilities

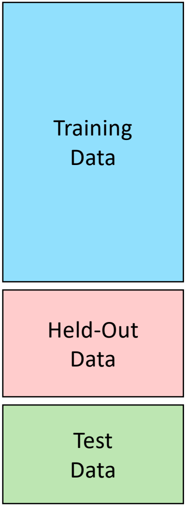
$$\hat{y} = \arg\max_k \; P(y_k) \prod_j P(x_j \mid y_k)$$

becomes:

$$\hat{y} = \arg\max_k \; \left[\log P(y_k) + \sum_j \log P(x_j \mid y_k) \right]$$

Naïve Bayes Learning

- Learn parameters from training data
  $P(Y), P(X_i|Y), P(X_j|Y), \ldots$
- Tune hyper-parameters on hold-out (validation) data
  For example, select smoothing hyperparameter m
- Choose best value, train final model on train+hold-out, evaluate final model on test data



Naïve Bayes Summary - Positives

- Very quick, scales to very large problems
- Simple to “train”, one pass through data to estimate probabilities
- Works very well despite strong assumption of conditional independence
  - there are some distributions too extreme and will fail, e.g., XOR
- Conditional independence assumption make Naïve Bayes good for high dimensional data
  - often not enough data for high dimensional problems without strong assumptions
  - may not estimate probabilities correct, but often makes correct decisions
- Robust to isolated noisy samples
- Handles missing values - drop samples

Naïve Bayes Summary - Negatives

- If features are not conditionally independent, introducing bias into classifier
- For continuous features,
    - binning loses information from the data, or
    - assumes Gaussian distribution, which may not be true
- Naïve Bayes does not do well or as well as other methods when:
    - there are repeated attributes
    - there is a lot of data and few attributes (other methods may have advantage)
    - the attributes are not equally important