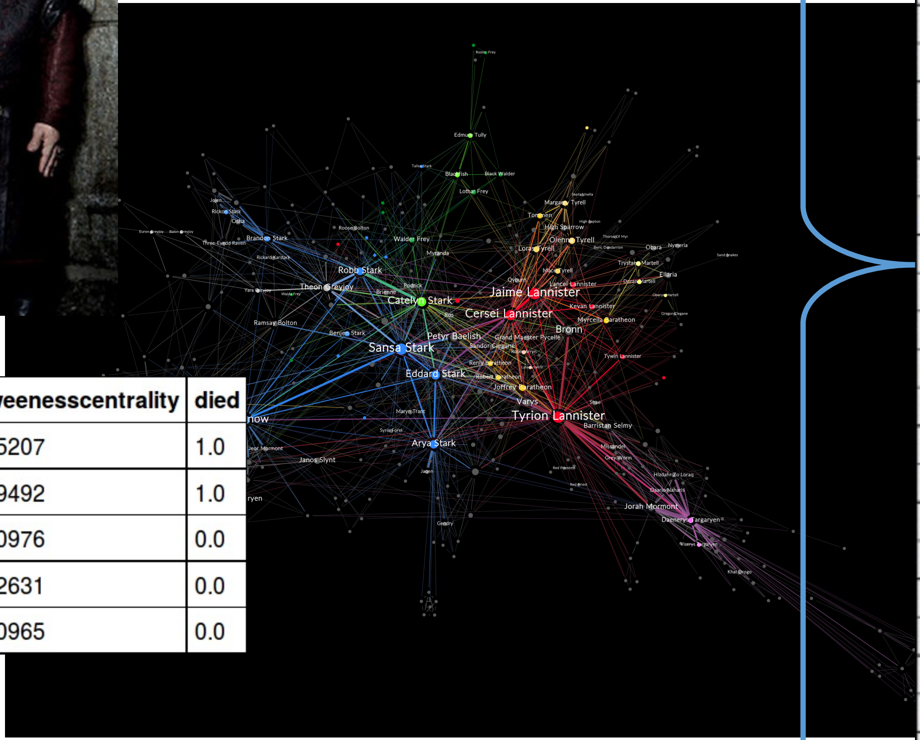

WON'T YOU BE
MY K- NEAREST NEIGHBOR?

https://networkdatascience.ceu.edu/article/2017-07-08/network-science-predicts-who-dies-next-game-thrones

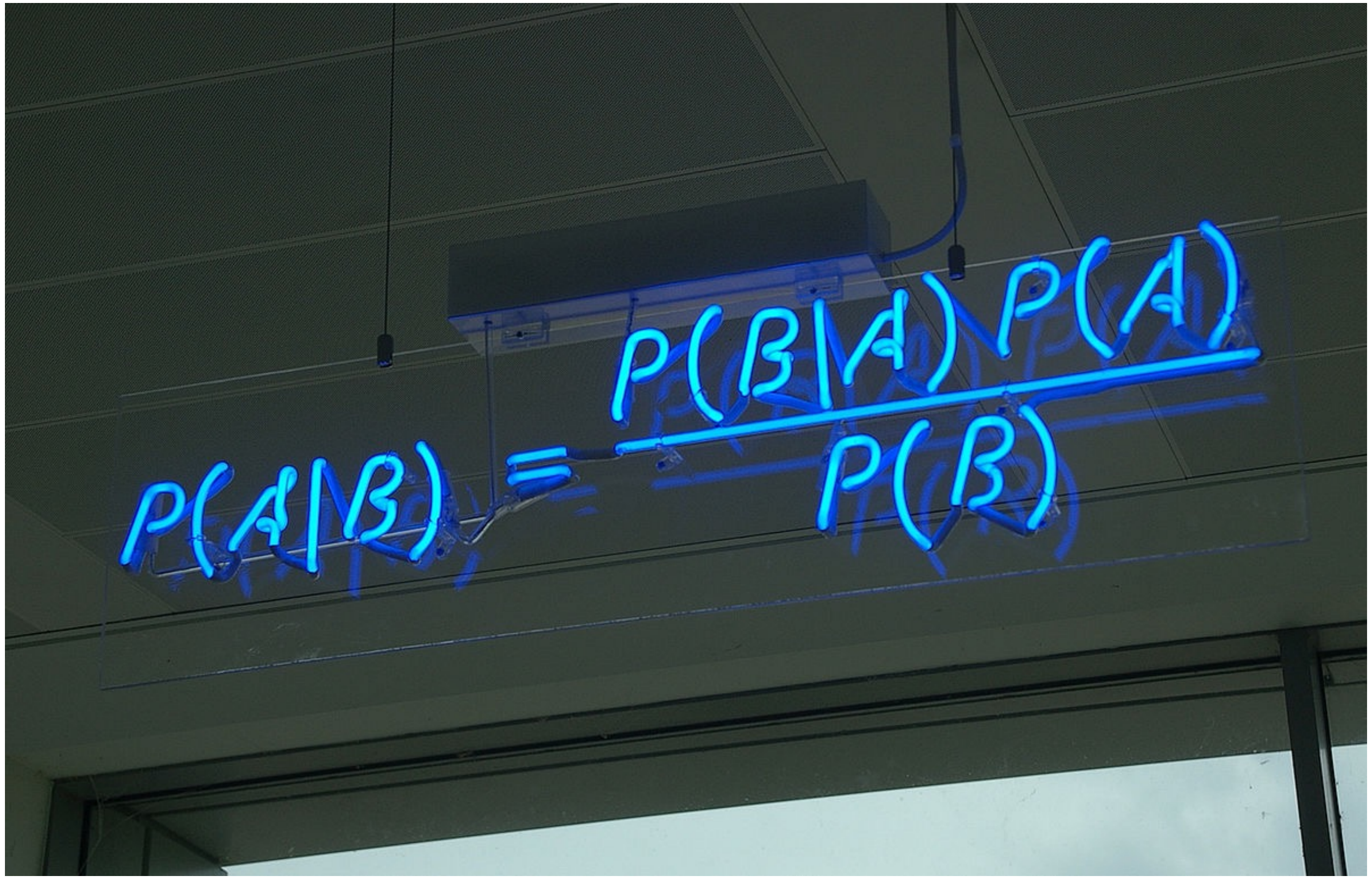


	degree	weighted degree	pageranks	clustering	eigencentrality	closnesscentrality	betweenesscentrality	died
Joffrey Baratheon	34	73	0.010648	0.567251	0.516207	0.510730	0.005207	1.0
Benjen Stark	23	34	0.007571	0.651515	0.411720	0.508547	0.009492	1.0
Theon Greyjoy	43	155	0.013937	0.324111	0.507033	0.548387	0.030976	0.0
Jon Snow	74	374	0.024013	0.240216	0.698632	0.595000	0.122631	0.0
Melisandre	15	59	0.005920	0.535714	0.106703	0.399329	0.000965	0.0

	Probability	Error
Tyene	0.95	0.04
Daenerys Targaryen	0.91	0.05
Grey Worm	0.90	0.05
Robin Arryn	0.90	0.06
Podrick	0.88	0.06
Edmure Tully	0.88	0.06
Greatjon Umber	0.88	0.06
Jaqen	0.87	0.06
Bronn	0.85	0.09
Arya Stark	0.84	0.07
Sandor Clegane	0.83	0.07
Brienne	0.81	0.07
Melisandre	0.80	0.07
Olenna Tyrell	0.79	0.09
Yara Greyjoy	0.74	0.07
Daario Naharis	0.72	0.09
Jaime Lannister	0.67	0.11
Gendry	0.65	0.13
Brandon Stark	0.64	0.12
Sansa Stark	0.59	0.11
Varys	0.56	0.12
Tyrion Lannister	0.52	0.12
Petyr Baelish	0.50	0.14
Qyburn	0.50	0.11
Jon Snow	0.45	0.10
Dolorous Edd	0.45	0.11
Gilly	0.37	0.12
Samwell Tarly	0.35	0.14
Cersei Lannister	0.35	0.16
Davos	0.32	0.08
Missandei	0.28	0.10
Obara	0.25	0.08
Tormund	0.14	0.08
Jorah Mormont	0.05	0.06
Theon Greyjoy	0.05	0.06

“Game of Thrones is a complex world in which social position and true friends seem to be quite important, so I quantified each character’s social interaction patterns using the tools of network science. I then predict their fate using machine learning methods.” –Milan Janosov

Naïve Bayes Classifier



Bayesian Classifiers

- Bayesian Classifiers take a probabilistic approach to classification
- Relationships between input features and class are expressed as probabilities

Naïve Bayes

- Uses Bayes Theorem and a "naïve" assumption
- Assume the input features are statistically independent from one another

Probability Review

Probability of A:

$$P(A) = \frac{\text{\# of ways for A to occur}}{\text{\# of possible outcomes}}$$

Joint probability of A & B occurring together, for **independent events**:

$$P(A,B) = P(A) * P(B)$$

Conditional probability:
Probability of event A occurring, given that B occurred

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

Joint probability of A & B occurring together, for **dependent events**:

$$P(A,B) = P(A) * P(B|A)$$

Marginal probability of A given all k possibilities for B:

$$P(A) = \sum_{i=1}^{k} P(A|b_i)P(b_i)$$

Bayes Theorem

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

Thomas Bayes (1702-1761)

Apply it! Example using Marginals

- Given the following (made up) example:
  - In Austin, 45% of registered voters are Democrats, 37% of registered voters are Republicans, and the remaining 18% are Independents. In the last election 35% of the Democrats, 62% of the Republications, and 58% of the Independents voted. A voter is chosen at random.

  - What fraction of registered voters voted in the election?

  - What is the probability that someone who voted is a Republican?

Q2.2a

What fraction of registered voters voted in the election?

Let:

R represent the randomly chosen registered voter is a Republican

D represent the randomly chosen registered voter is a Democrat

I represent the randomly chosen registered voter is an Independent

V represent the randomly chosen registered voter actually voted

$P(D) = 0.45$

$P(R) = 0.37$

$P(I) = 0.18$

$P(V|D) = 0.35$

$P(V|R) = 0.62$

$P(V|I) = 0.58$

$P(V) = \sum_{party} P(V|party)P(party) = P(D) * P(V|D) + P(R) * P(V|R) + P(I) * P(V|I) = 0.45 * 0.35 + 0.37 * 0.62 + 0.18 * 0.58 = 0.4913$
$= 49.13\%$

Q2.2b

What is the probability that someone who voted is a Republican?

$$P(R|V) = \frac{P(R)*P(V|R)}{P(V)} = \frac{0.37*0.62}{0.4913} = 0.467 = 46.7\%$$

Classification with Bayes Theorem

Given features $X = \{X1, X2, X3, \ldots Xn\}$, predict class C.

Do this by finding the class C that maximizes $P(C \mid X)$.

$$\left.\begin{array}{l} P(C1 \mid X) \\ P(C2 \mid X) \\ P(C3 \mid X) \\ \ldots \\ P(Ck \mid X) \end{array}\right\} \quad \text{Max => Class Label}$$

Classification with Bayes Theorem

Class-conditional probability:
Probability of observing input features X, given that C is the class label

Prior probability:
Probability of class label being C, prior to observing anything

$$P(C\,|\,X) = \frac{P(X\,|\,C) * P(C)}{P(X)}$$

Posterior probability:
Probability of class label being C, after observing input features X

Probability of observing input features X, regardless of what the class label is (**probability of the evidence**)

Classification with Bayes Theorem

Class-conditional probability:
Use the "naïve" independence assumption to simplify (see next slide)

Prior probability:
Calculate the fraction of records with each class label in the training data

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Posterior probability

Constant that can be ignored! Does not depend on class, so is the same for all classes, therefore can be ignored.

Calculating Class-Conditional Probability

- Use the "naïve" independence assumption to simplify

$$
\begin{aligned}
& P(X \mid C) \\
&= P(X1, X2, X3, \ldots, Xn \mid C) \\
&= P(X1 \mid C) * P(X2 \mid C) * P(X3 \mid C) * \ldots * P(Xn \mid C)
\end{aligned}
$$

Example

$$P(C|X) \propto P(X|C) * P(C)$$

$$P(C|X) \propto P(C) \prod P(x_i|C))$$

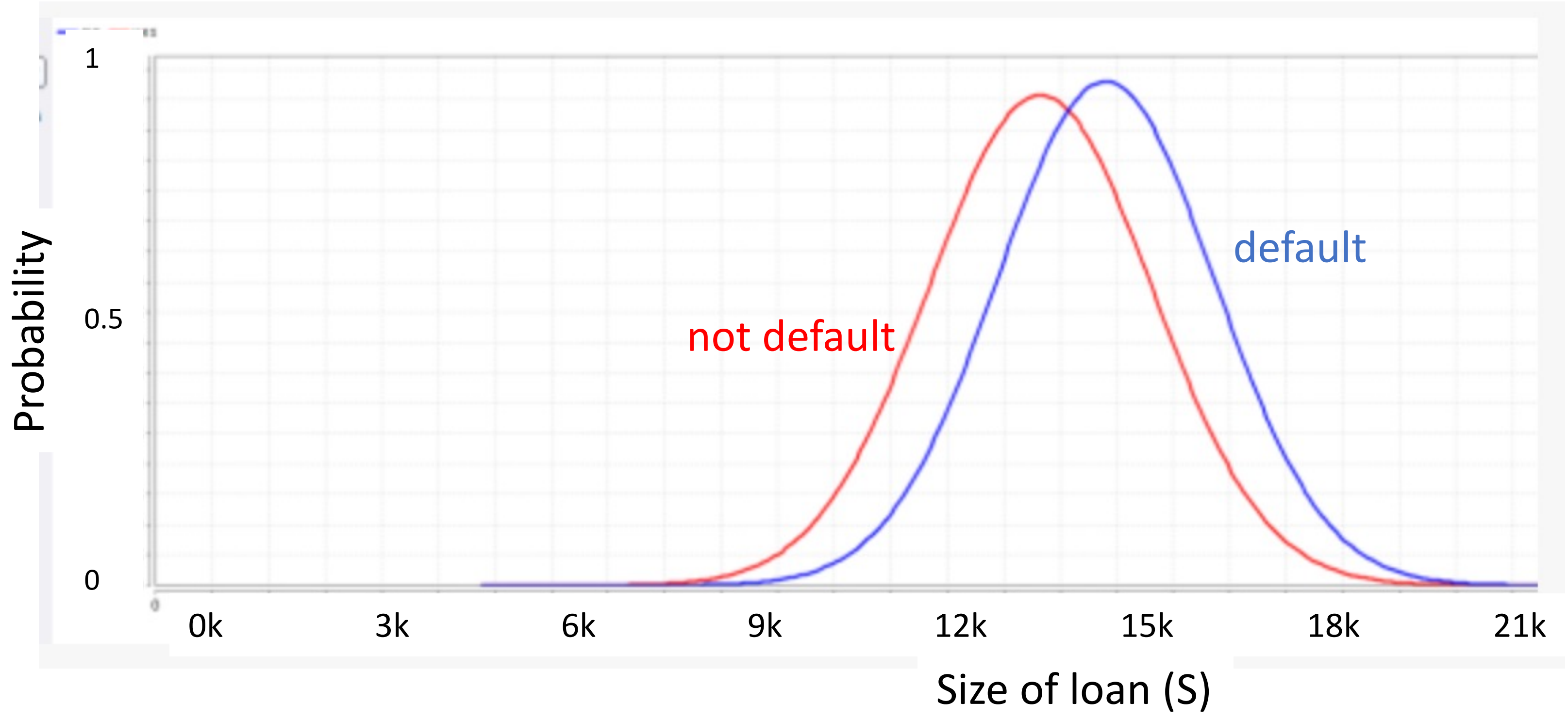
Classify new record:
T=9;   S=5,000;   H=own

Time with employer (T)	Size of loan (S)	Homeowner (H)	Default (D)
5	10,000	own	no
20	10,000	own	no
1	25,000	own	no
1	15,000	other	no
15	2,000	rent	no
6	12,000	own	no
1	5,000	rent	yes
12	8,000	rent	yes
3	10,000	own	yes
1	5,000	other	yes

T <= 10   S <= 10,000

Continuous Attributes

- Binning
- Probability Density Function



Estimate Probabilities from Data

- For continuous attributes:
  - Discretization: Partition the range into bins:
    - Replace continuous value with bin value
      - Attribute changed from continuous to ordinal
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (X_i, Y_i) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Laplace Smoothing

- If we try to classify a record that contains an attribute value we've never seen before, all of the probabilities zero out

Training Data

Color	Flavor	Weight	Class
Yellow	Sweet	2	Banana
Yellow	Sour	3	Apple
Red	Sweet	2	Apple
Red	Sour	4	Apple
Green	Sweet	3	Apple

Now, try to classify something that is Green…

Laplace Smoothing

- Use "Laplace smoothing" (or "additive smoothing") to avoid zero-ing out your equation
- For all class-conditional probabilities for this attribute:
  - Add 1 to the numerator
  - Add v to the denominator, where v is the total number of values this attribute can take
- You do this "add 1" anytime there is an attribute value that does not occur in every class.

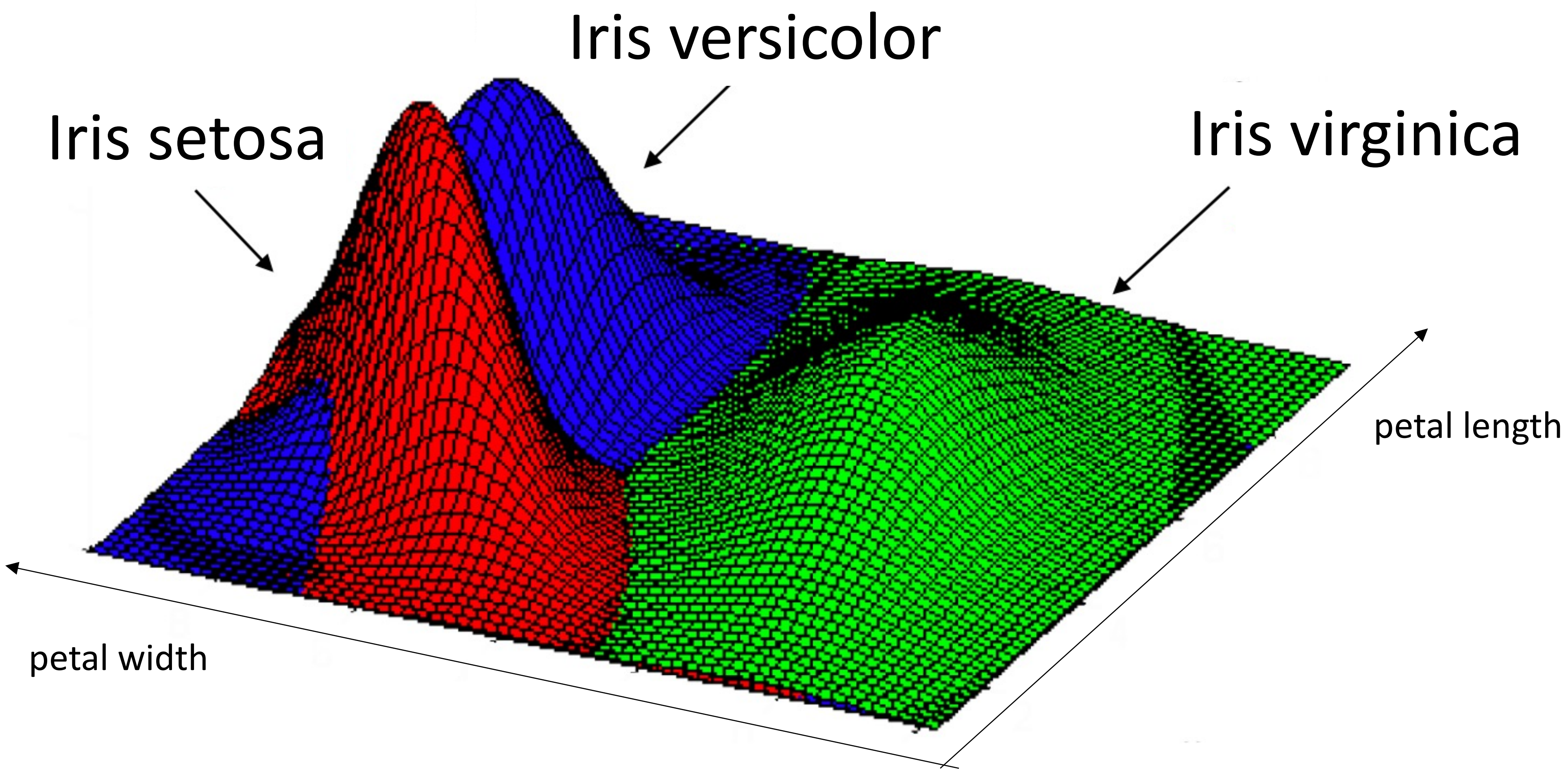
Training Data

Color	Flavor	Weight	Class
Yellow	Sweet	2	Banana
Yellow	Sour	3	Apple
Red	Sweet	2	Apple
Red	Sour	4	Apple
Green	Sweet	3	Apple

$$P(X_i = c|y) = \frac{n_c + 1}{n + v} \qquad P(banana|green) = \frac{0+1}{1+3} = 0.25$$

Where n is the number of training instances belonging to the class y, nc is the number of training instance with Xi = c and v is the total number of attribute values that Xi can take

Decision Boundaries / Probability Distributions



Characteristics of Naïve Bayes

- Fast and simple algorithm
  - Probabilities can all be calculated with a single scan of the data, then stored; iterative processing of the data is not necessary
- Scales well – avoids problems with the curse of dimensionality
- Independence assumption may not always hold true
  - In practice, Naïve Bayes still works quite well

- Handle missing values by ignoring the instance during probability estimate calculations

Use Naïve Bayes:
Use the Laplace estimate if conditional probability is zero

Given all the previous patients I've seen (below are their symptoms and diagnosis)...

Chills	Runny nose	Headache	Fever	Flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	Y	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Do I believe that a patient with the following symptoms has the flu?

Y	N	Mild	Y	?