Applied Artificial Intelligence

Session 7: Probability and Statistics for AI and Machine Learning II

Fall 2018

NC State University

Instructor: Dr. Behnam Kia

Course Website: https://appliedai.wordpress.ncsu.edu/

Random Experiment

Sample Space: $\Omega = \{Head, Tail\}$

Random Variable

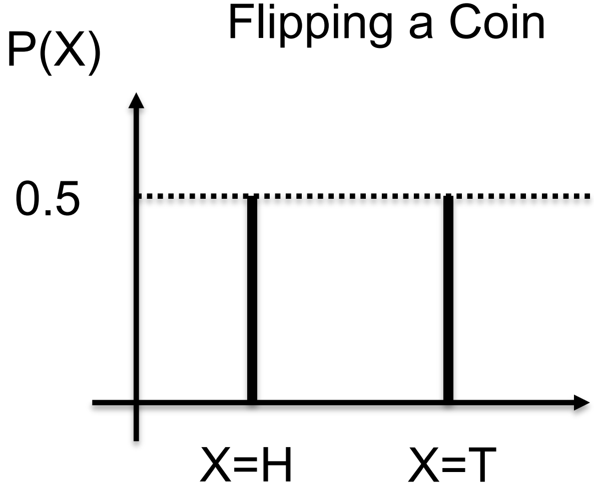
- A random variable is a variable that takes on different values based on outcomes of a random experiment.

- Random variable can be discrete or continuous.

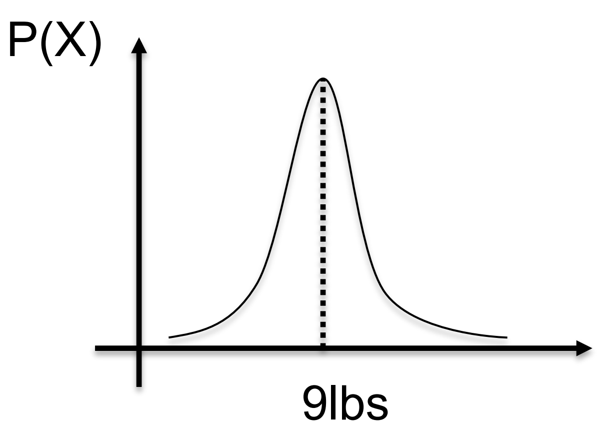
Probability Distribution Function

- Probability distribution function is a description of how likely a random variable or a set of variables is to take on each of its possible states.

Flipping a Coin



Cat's weight

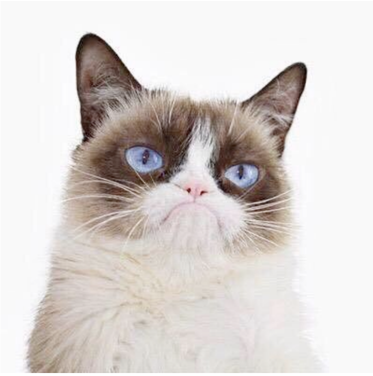


Hypothetical Scenario

- We are installing cameras and sensors in a neighborhood to record the presence of feral animals.
- There are two types of animals in the neighborhood; Dogs and Cats.
- We like to design an automatic system to determine whether the recorded animal is a cat or a dog.

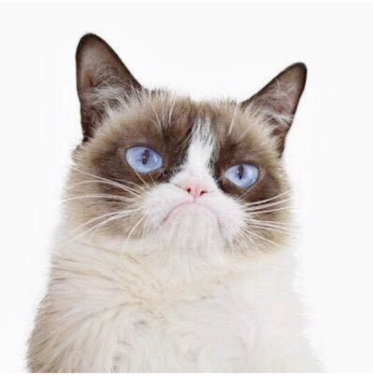
Hypothetical Scenario

- Ratio of Cats to Dogs is 1 to 3. (75% are dogs, 25% cats)



Hypothetical Scenario

- Ratio of Cats to Dogs is 1 to 3. (75% are dogs, 25% cats)
- Making a decision without looking at the sensor readings.



?

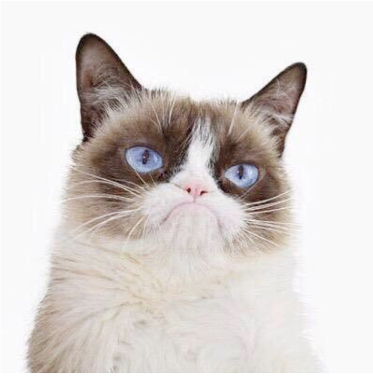
Hypothetical Scenario

- Ratio of Cats to Dogs is 1 to 3. (75% are dogs, 25% cats)

- Making a decision without looking at the sensor readings.

P(Cat)=0.25

P(dog)=0.75



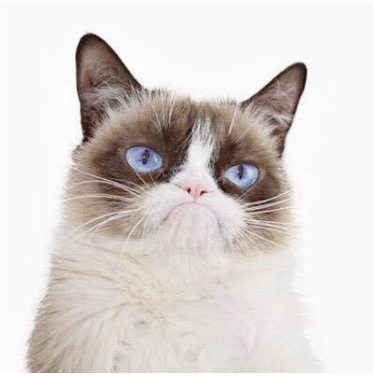
?

Hypothetical Scenario

- Ratio of Cats to Dogs is 1 to 3. (75% are dogs, 25% cats)
- Making a decision without looking at the sensor readings.
- **What is the expected error rate**?

P(Cat)=0.25

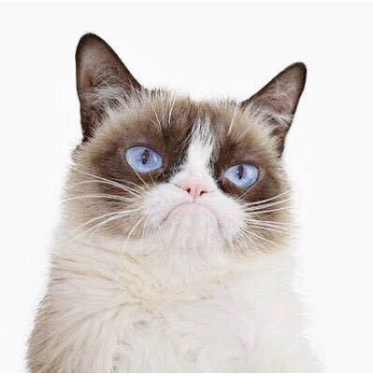
P(dog)=0.75



Hypothetical Scenario

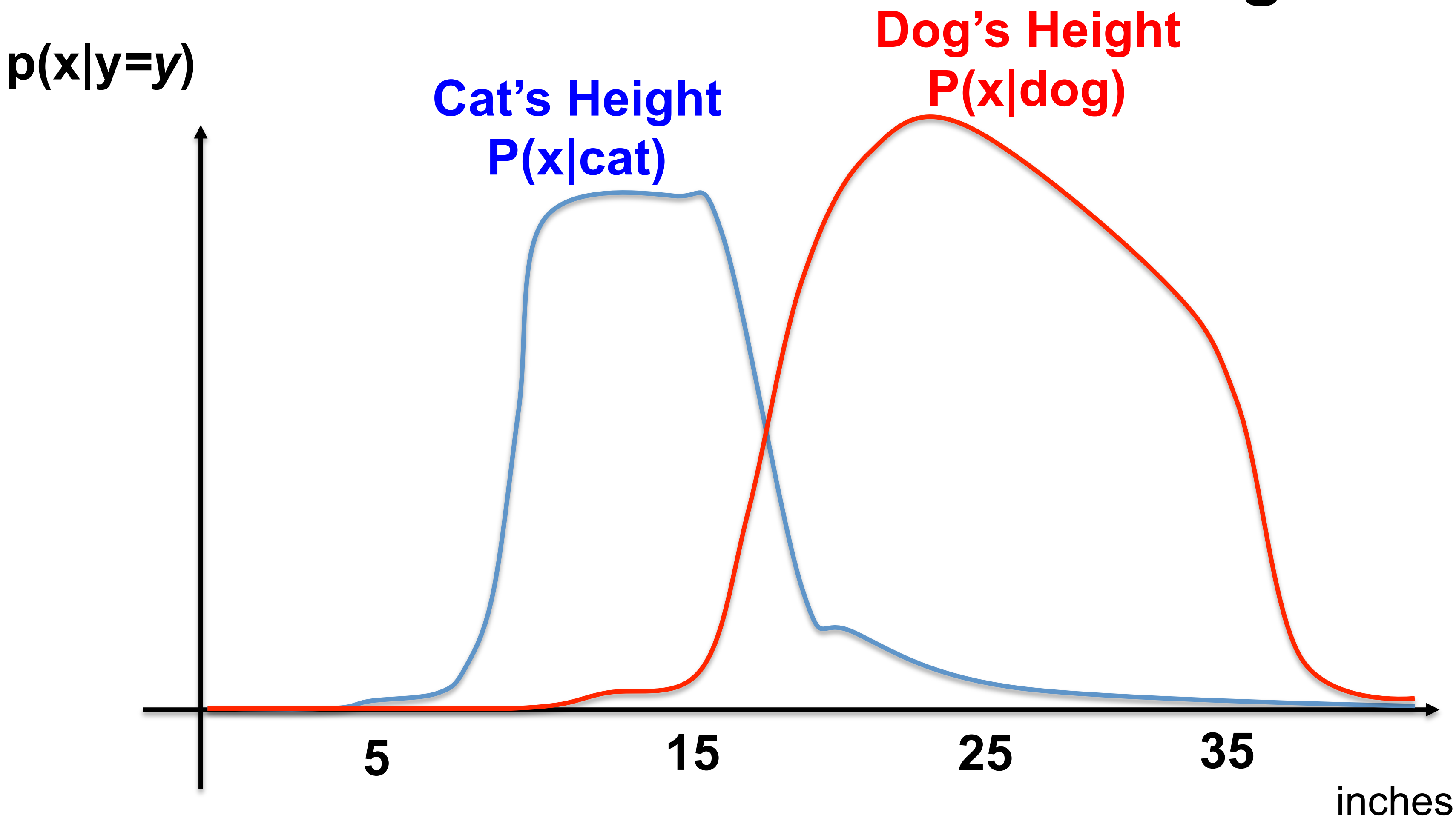
- Ratio of Cats to Dogs is 1 to 3. (75% are dogs, 25% cats)
- And we have access to some data, the height of the animal being recorded.

P(Cat)=0.25

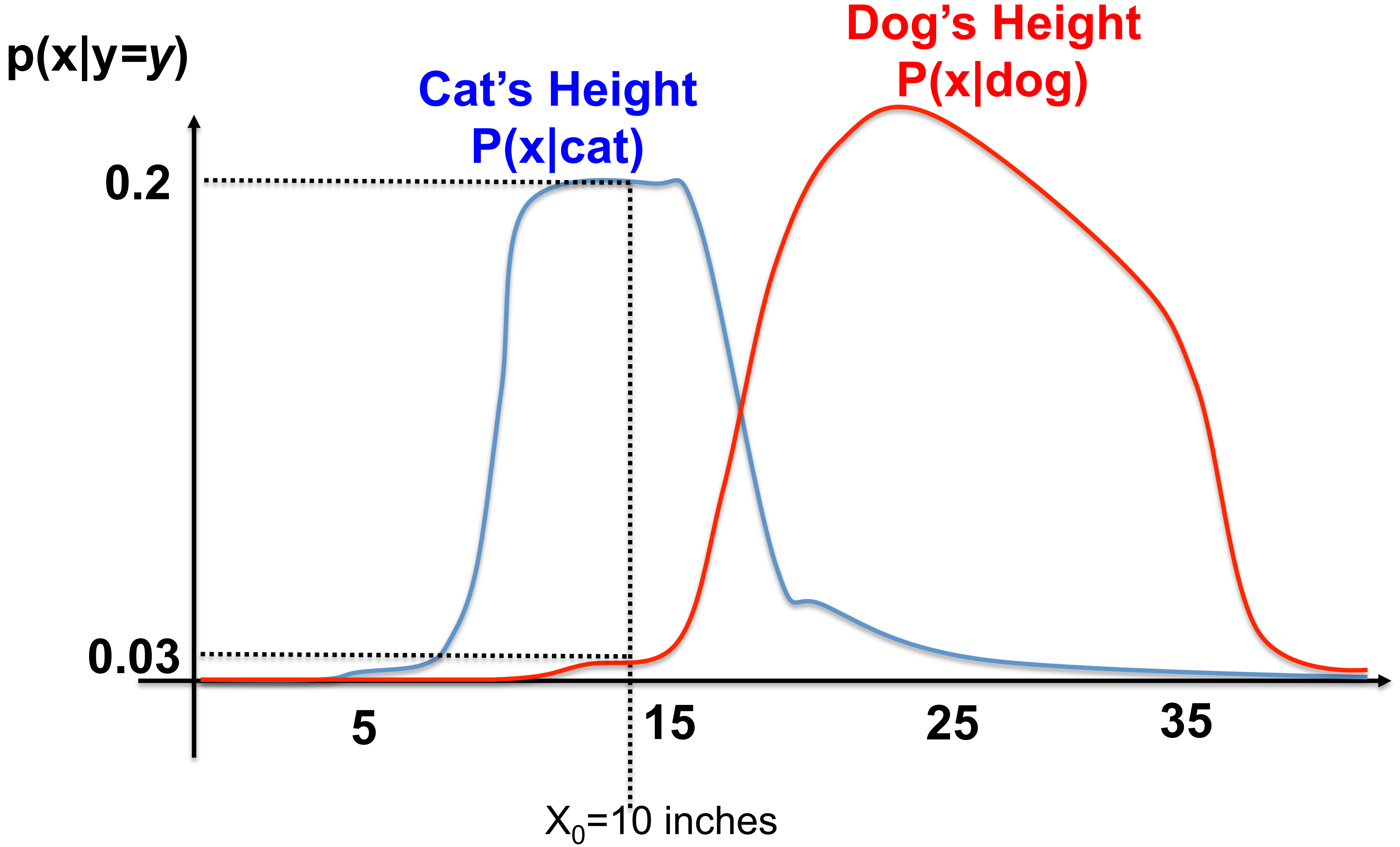


P(dog)=0.75

Class-Conditional Probability Distribution Function for Height



Class-Conditional Probability Distribution Function for Height



Bayes' Rule

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

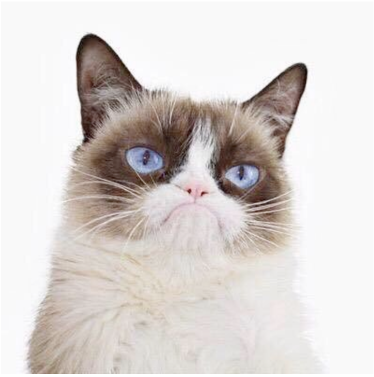
$$posterior = \frac{likelihood \times prior}{evidence}$$

Bayes' Rule

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

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

$$p(y = cat \mid x = 10) = \frac{p(x = 10 \mid y = cat)\, p(cat)}{p(x = 10)}$$



$$p(y = dog \mid x = 10) = \frac{p(x = 10 \mid y = dog)\, p(dog)}{p(x = 10)}$$

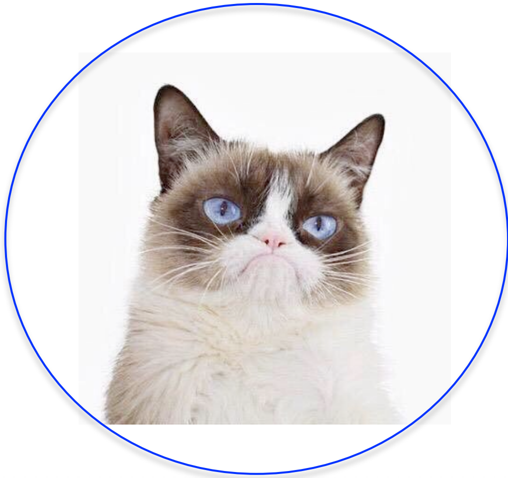
Bayes' Rule

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

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

$$p(y = cat \mid x = 10) = \frac{p(x = 10 \mid y = cat)\, p(cat)}{p(x = 10)}$$

$$p(y = cat \mid x = 10) = \frac{0.2 \times 0.25}{p(x = 10)} = \frac{0.05}{p(x = 10)}$$



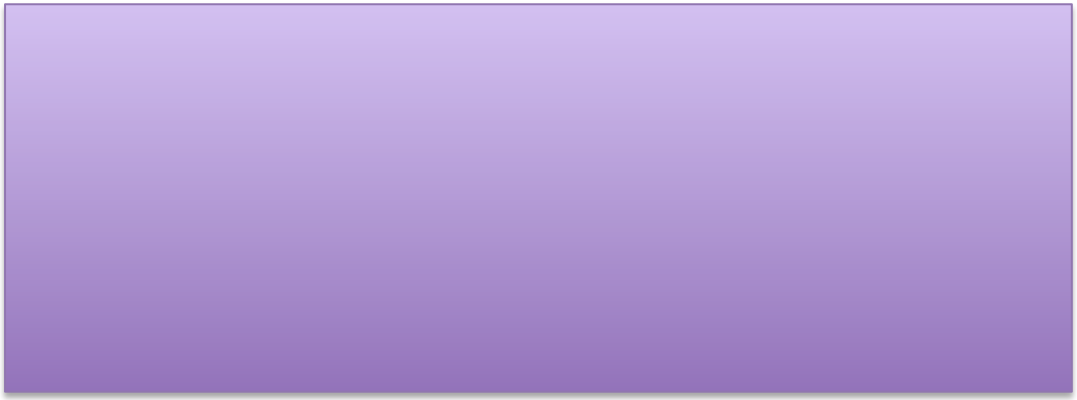
$$p(y = dog \mid x = 10) = \frac{p(x = 10 \mid y = dog)\, p(dog)}{p(x = 10)}$$

$$p(y = dog \mid x = 10) = \frac{0.03 \times 0.75}{p(x = 10)} = \frac{0.0225}{p(x = 10)}$$

p(evidence) Calculated Based on Law of Total Probability

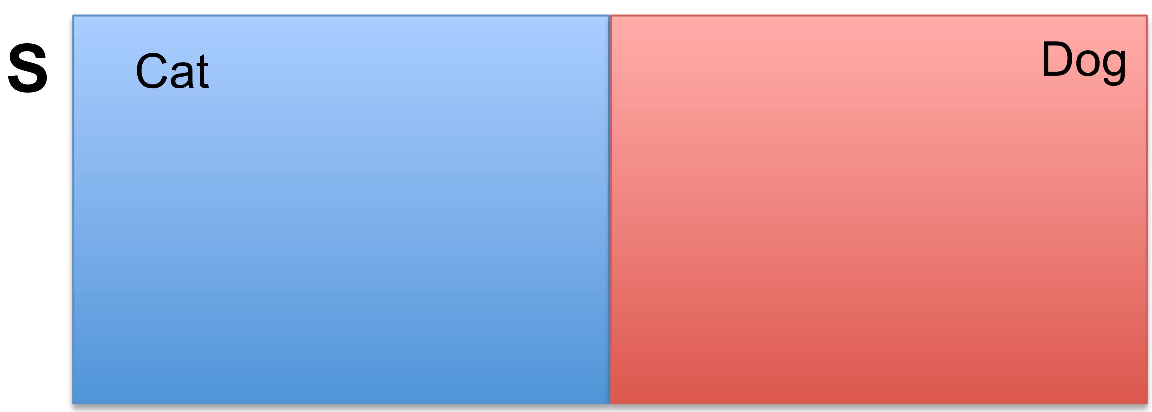
$$p(x = 10) = ?$$

S



p(evidence)
Calculated Based on Law of Total Probability

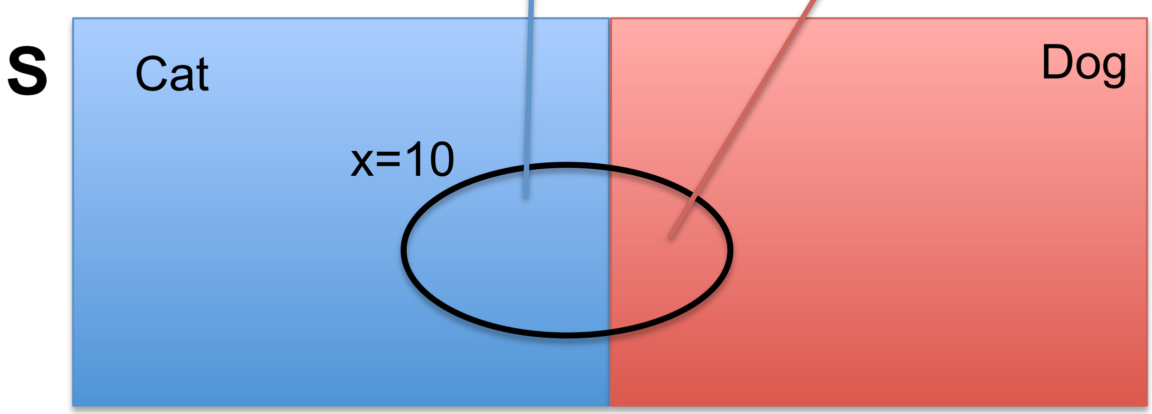
$$p(x = 10) = ?$$



$$S = Cat \bigcup Dog$$

p(evidence) Calculated Based on Law of Total Probability

$$p(x=10) = p(x=10 \cap y = Cat) + p(x=10 \cap y = Dog)$$



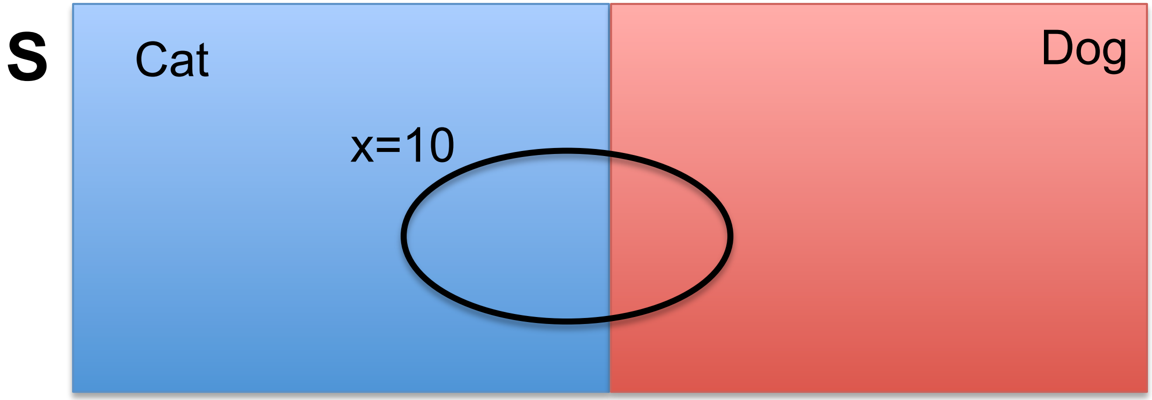
p(evidence) Calculated Based on Law of Total Probability

$$p(x=10) = p(x=10 \cap y=Cat) + p(x=10 \cap y=Dog)$$

$$p(x=10) = p(x=10 \mid y=cat)\,p(y=cat) + p(x=10 \mid y=dog)\,p(y=dog)$$

$$p(x=10) = 0.2 \times 0.25 + 0.03 \times 0.75$$

$$= 0.0725$$

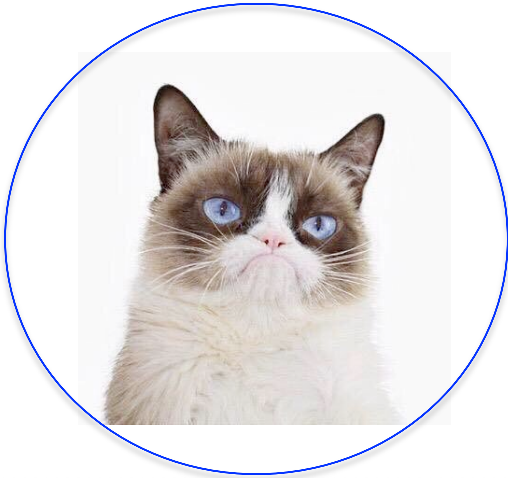


Bayes' Rule

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

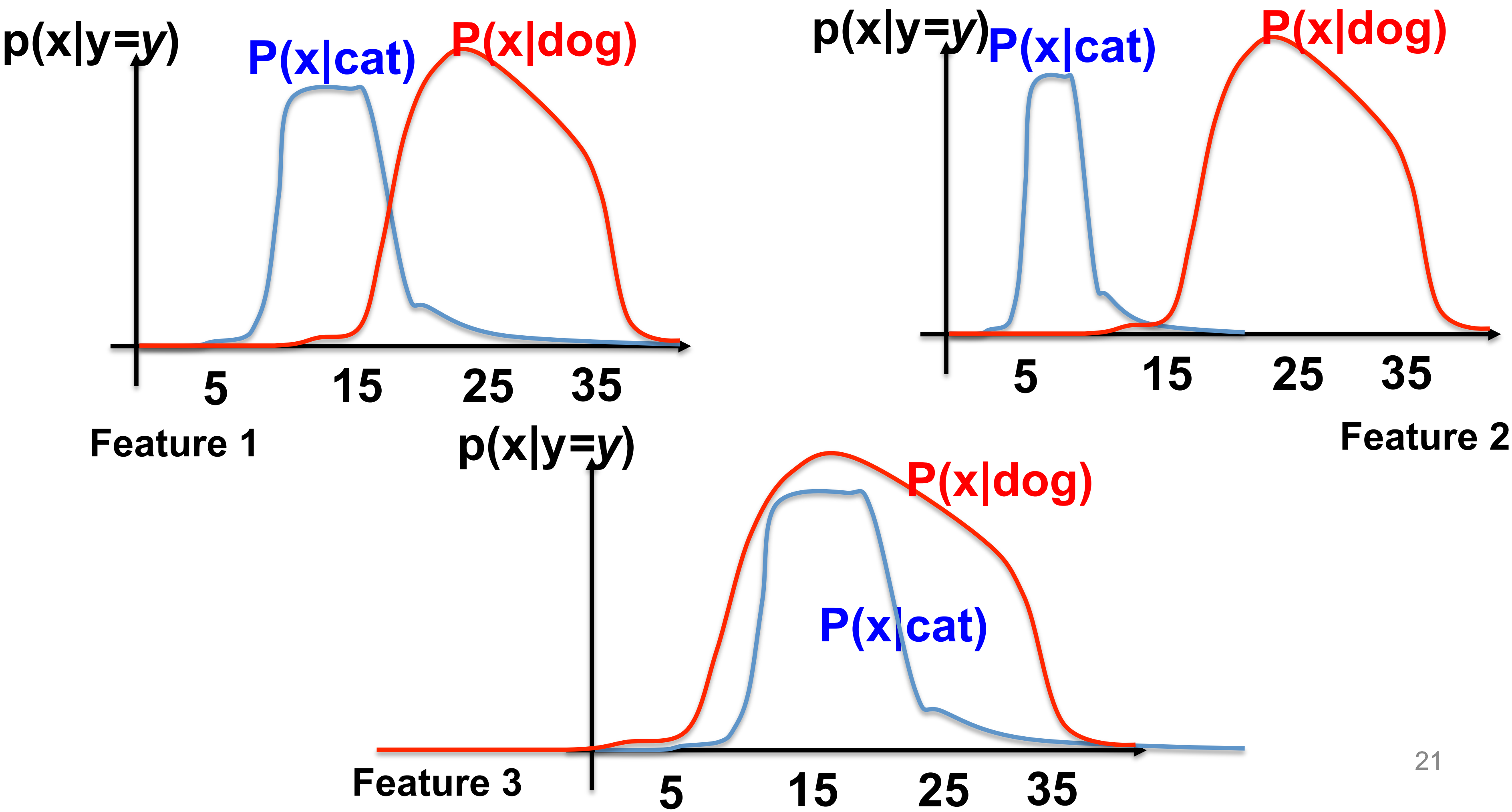
$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

$$p(y = cat \mid x = 10) = \frac{0.2 \times 0.25}{p(x = 10)} = \frac{0.05}{p(x = 10)} = 0.69$$

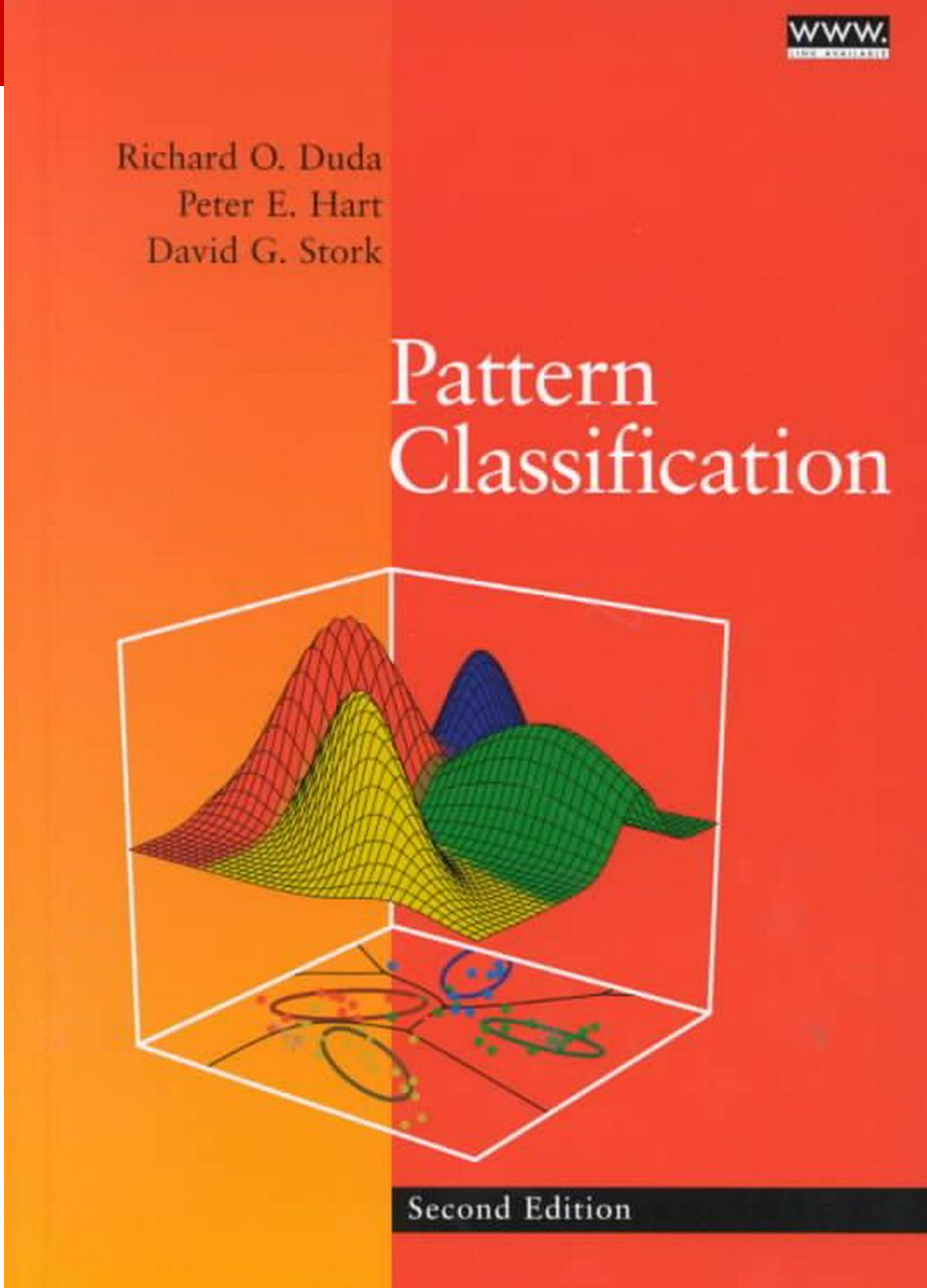


$$p(y = dog \mid x = 10) = \frac{0.03 \times 0.75}{p(x = 10)} = \frac{0.0225}{p(x = 10)} = 0.31$$

Which Feature *x* Would You Choose?

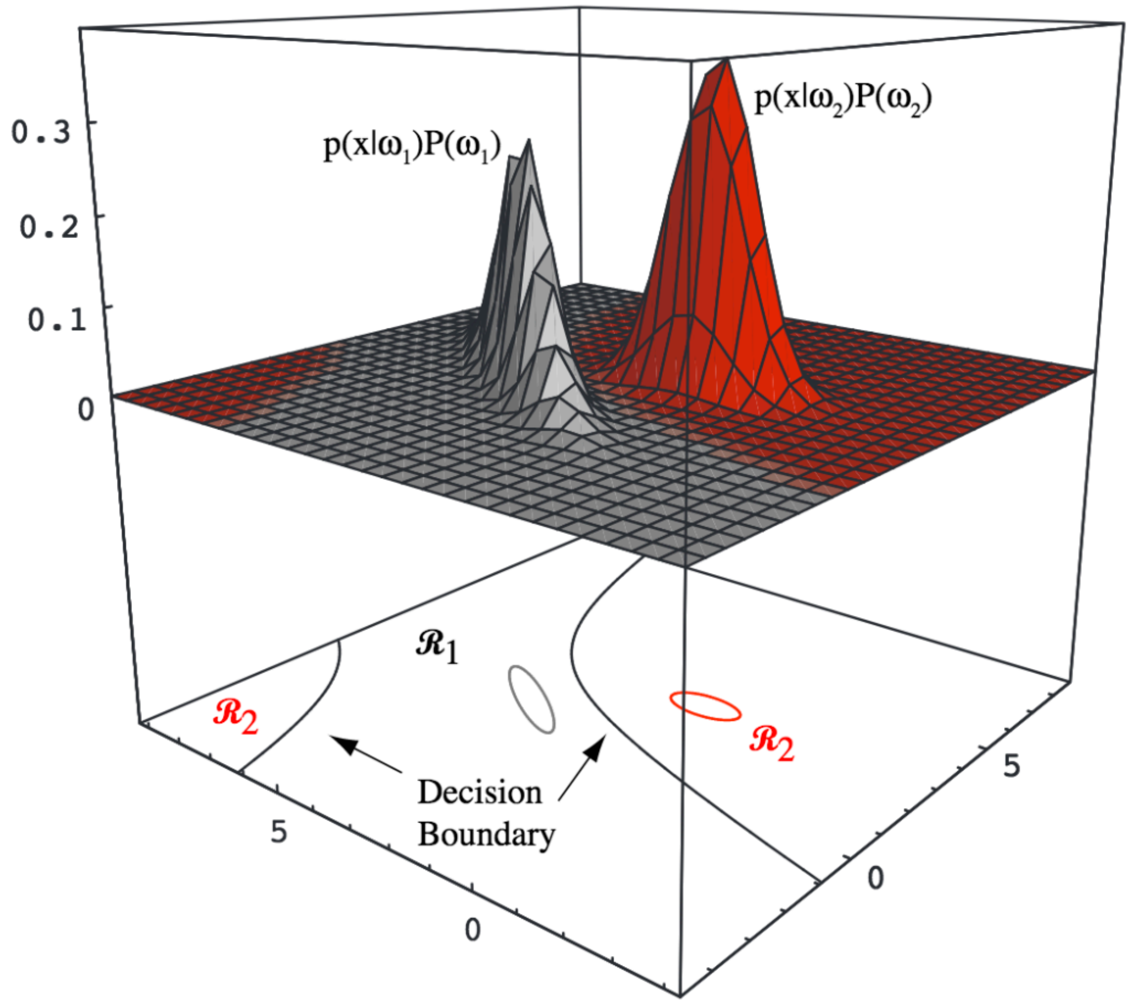


Same Concepts, But in 2D



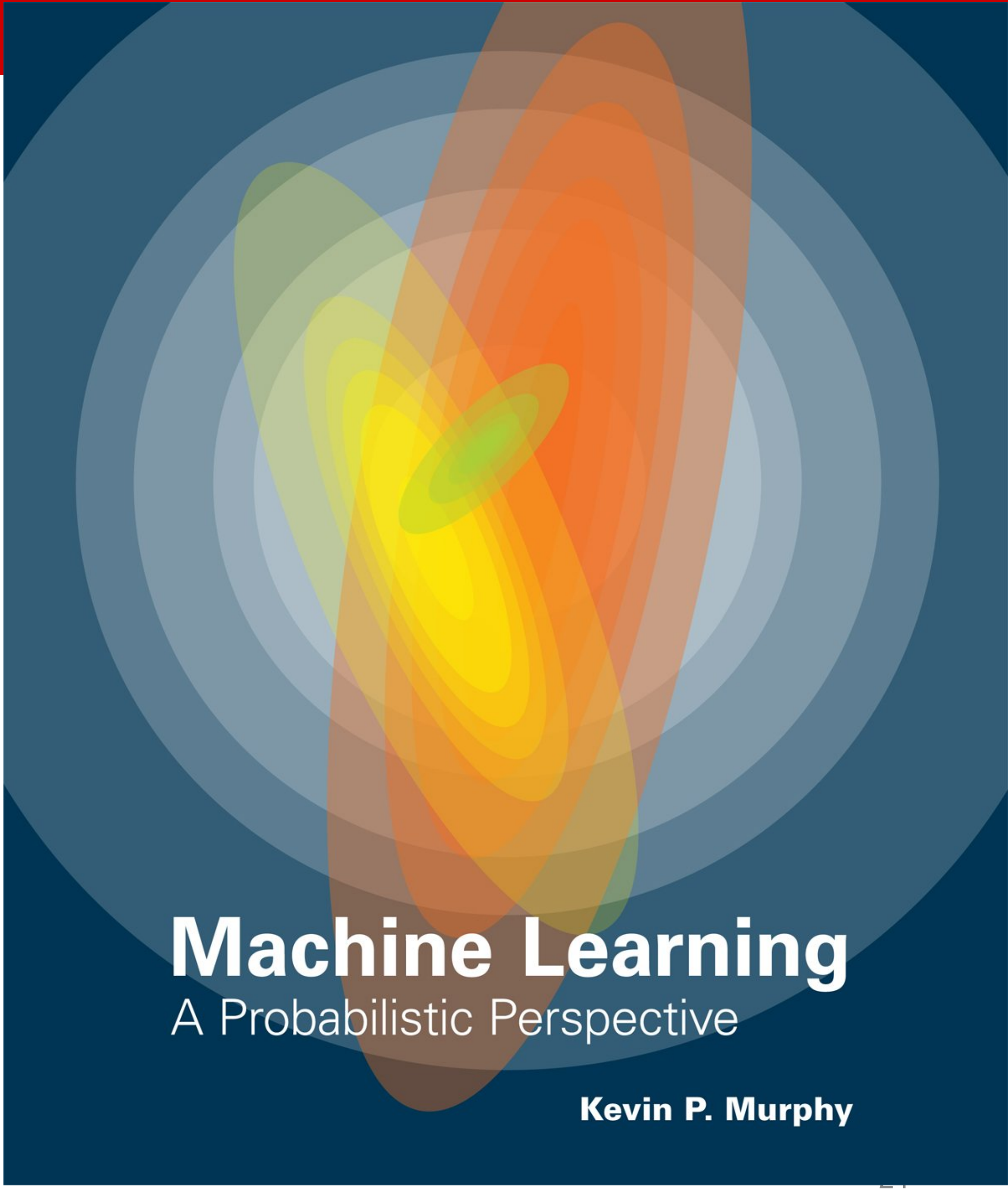
Pattern Classification,
Duda, Hart, and Stork

Same Concepts But in 2D



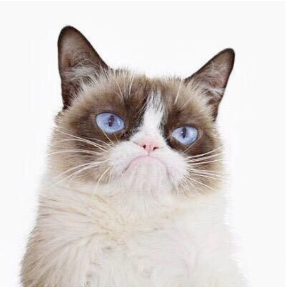
Pattern Classification, Duda, Hart, and Stork

Same Concepts But in 2D



Machine Learning
A Probabilistic Perspective
Kevin Murphy

Classic Machine Learning



Feature Extraction

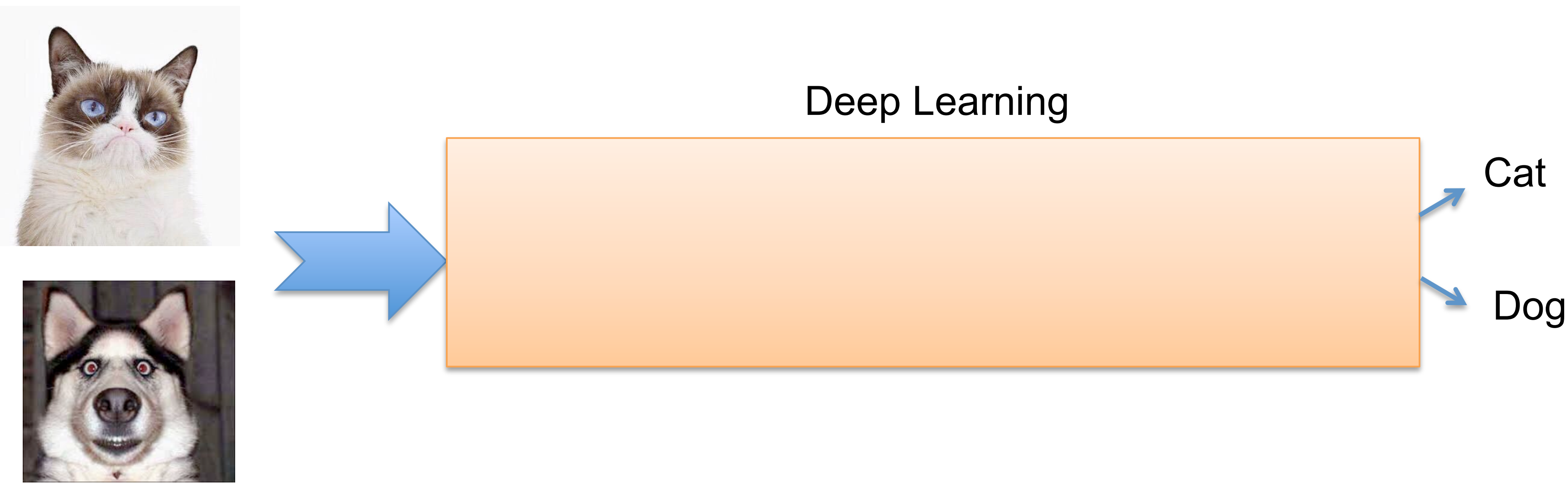
Height
Weight
Ratio of head to body
…

Machine Learning

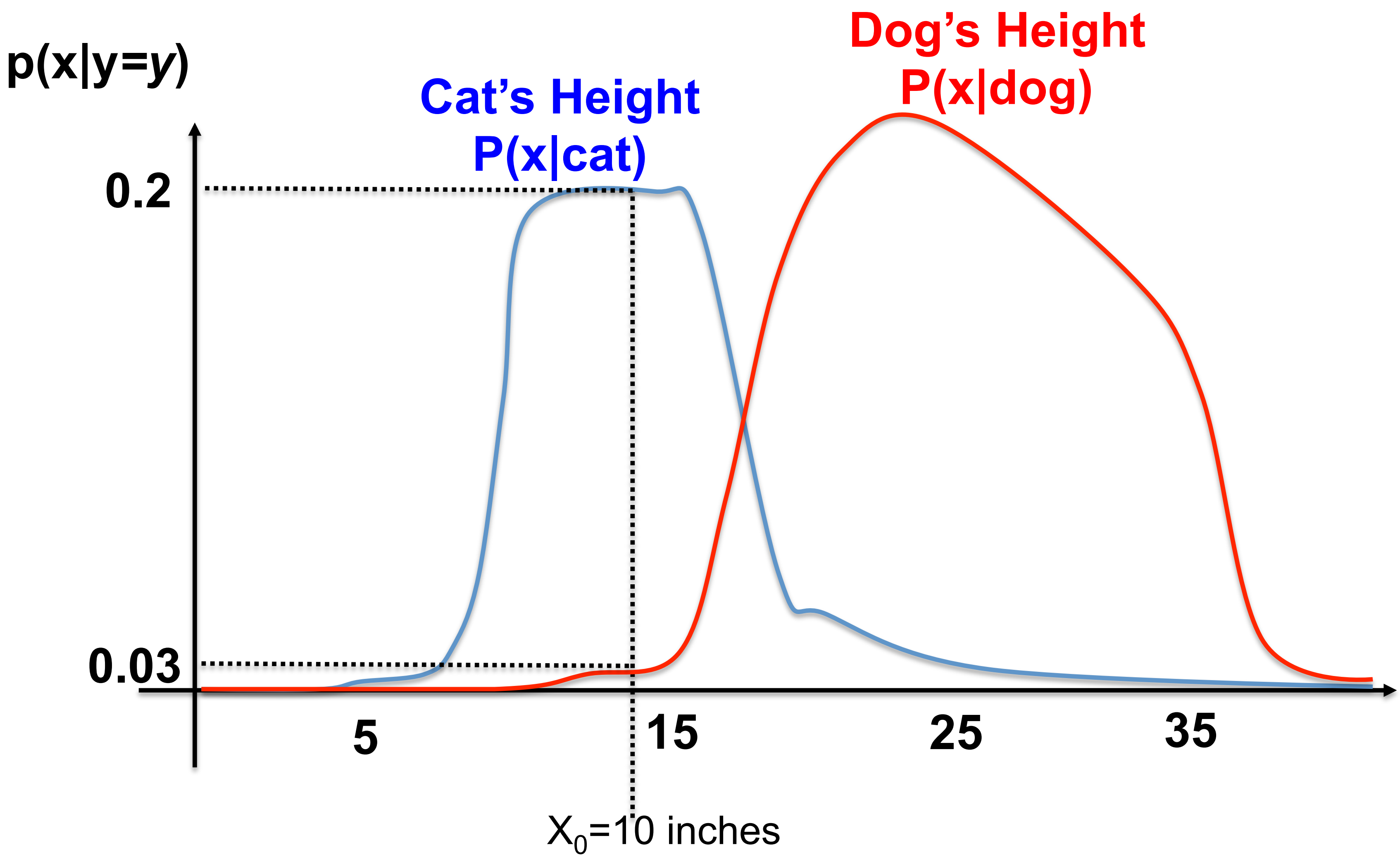
Cat

Dog

Deep Learning



How to Get Class-Conditional Probability Distribution Functions?



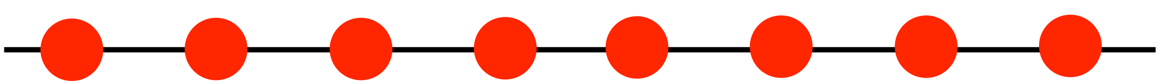
Machine Learning with Many Features

$$p(y \mid X) = \frac{p(X \mid y)\,p(y)}{p(X)}$$

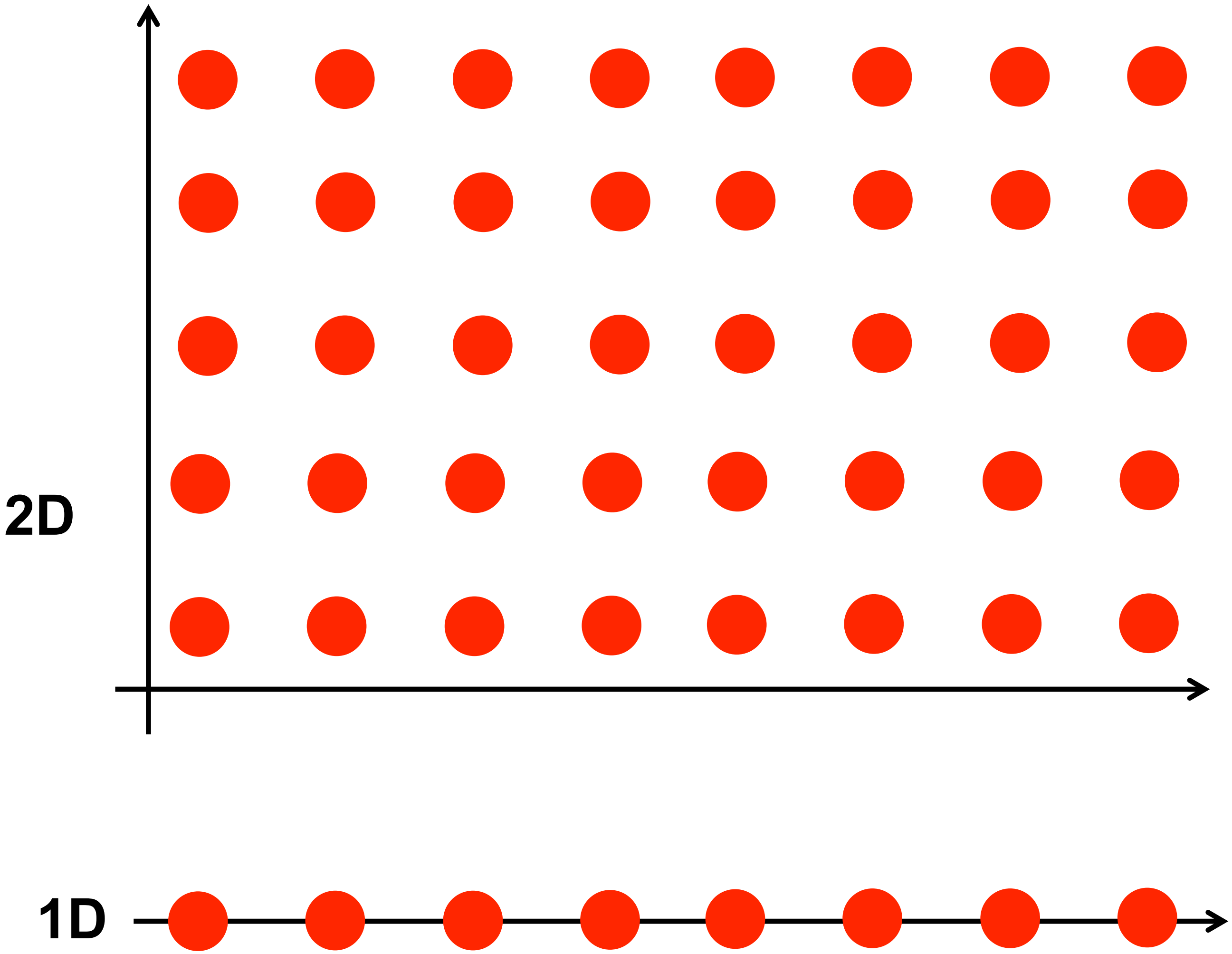
$$X = (x_1, x_2, \ldots, x_k)$$

The Curse of Dimensionality

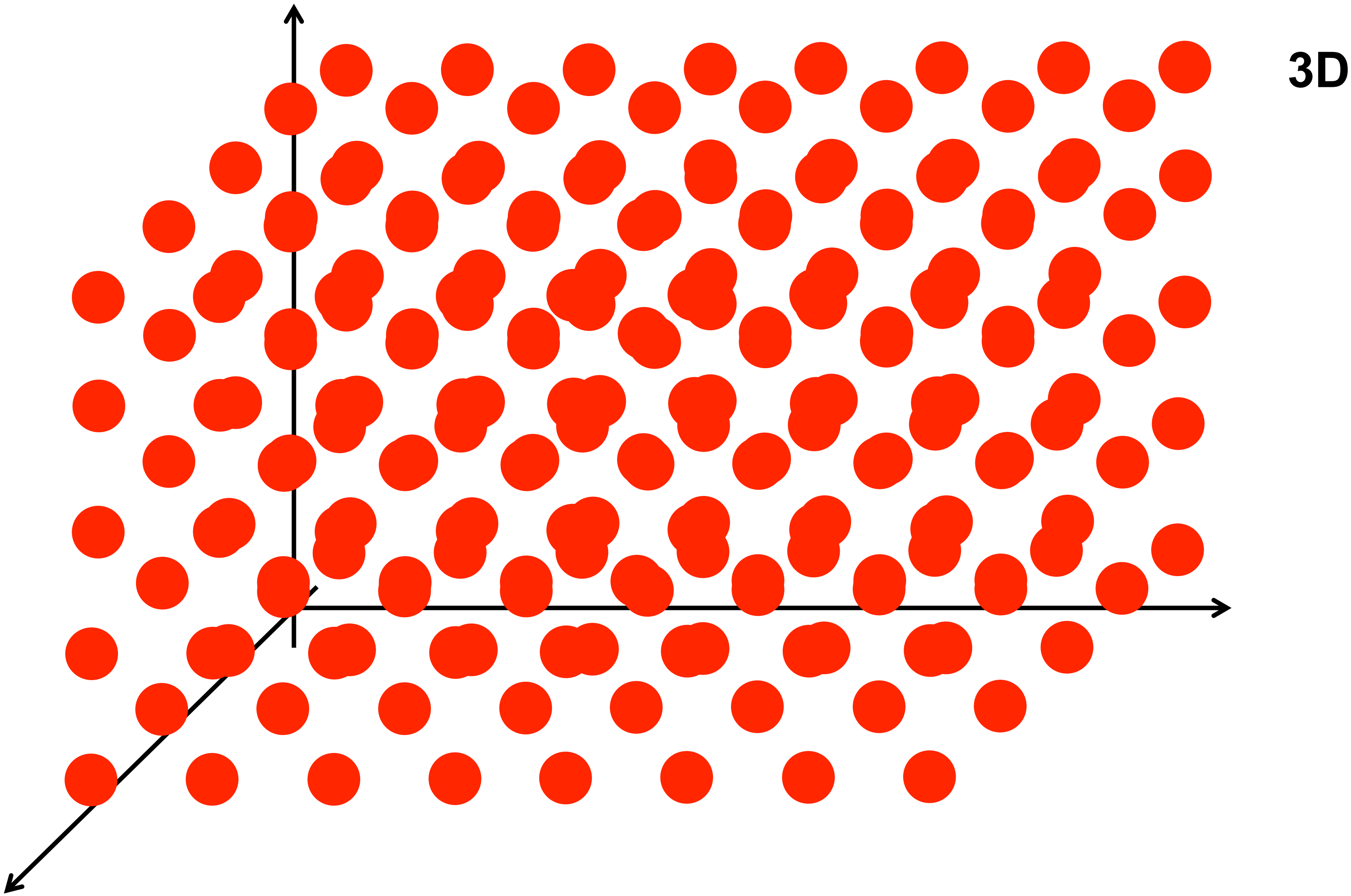
The Curse of Dimensionality

1D 

The Curse of Dimensionality



The Curse of Dimensionality



The Curse of Dimensionality

Naïve Bayes Rule

Bayes Rule

$$p(y \mid x_1, x_2, \ldots, x_k) = \frac{p(x_1, x_2, \ldots, x_k \mid y)\, p(y)}{p(X)}$$

$$\approx \frac{p(x_1 \mid y)\, p(x_2 \mid y) \ldots p(x_k \mid y)\, p(y)}{p(X)}$$

Naïve Bayes Rule

We have assumed that features x_1, x_2,…, and x_k are conditionally independent given y

$$p(x_1, x_2, \ldots, x_k \mid y) \approx p(x_1 \mid y)\, p(x_2 \mid y) \ldots p(x_k \mid y)$$

Document Classification Using Naïve Bayes Rule

(Homework III)

Document Classification

- Imagine we are given a document. We would like to classify it. For example:
  - An email is spam or ham?
  - A review is positive or negative? (sentiment analysis)
  - The subject of the document is Math, Physics, or Chemistry?
  - Authorship identification
  - …

Document Classification
A review is positive or negative? (sentiment analysis)

Hahaha I knew it.
[redacted] 13 September 2018

Straight trash. Why bother making a predator movie if you are just going to ignore all the stories? Heck they even ignored all we have learned from the movies. It seems anyone can defeat a predator nowadays. Back in the day, it took Arnie and a whole squad of macho man to try to take on just one. Now any ole joe smoe can go toe to toe with the predator. Stupid.

Classifier

Positive ? Negative

Document Classification
A review is positive or negative? (sentiment analysis)

- Rule based approach:
  - If the review contains:

    "What an awful movie" OR

    "I need my money back!" OR

    "I wish I had got sick so I couldn't end up going to watch this movie!"

  Then it is <u>Negative</u>!

Document Classification A review is positive or negative? (sentiment analysis)

- Rule based approach:
  - If the review contains:

    "What a fun movie" OR

    "I am going to watch it again!" OR

    "This movie is the best thing that has happened to human race!"

  Then it is Positive!

Document Classification
A review is positive or negative? (sentiment analysis)

- Machine learning approach:
  - A training set of m labeled documents (R_1,c_1), (R_2,c_2), …, (R_m,c_m) $c_1, c_2, \ldots, c_m \in \{Positive, Negative\}$
  - Train a classifier that automatically assigns an unlabeled review to its correct class.
  - Many different machine learning techniques for this problem; here we use Naïve Bayes' Rule.

Document Representation

- Different representations.
- Here we use "*the bag of words" representation*. Order of the words doesn't matter, just the words and how many times they occur in the text.

The Bag of Word Representation

- Review to be classified: “this was a good movie. This was the best movie of the series.”

The Bag of Words Representation of the Review

this	2
was	2
a	1
good	1
movie	2
the	1
best	1
of	1
the	1
series	1

Document Classification

c is the class, d is the document we like to classify

$$p(c \mid d) = \frac{p(d \mid c)\, p(c)}{p(d)}$$

Document Classification

c is the class, d is the document we like to classify

$$c_{MAP} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} p(c \mid d)$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \frac{p(d \mid c)\, p(c)}{p(d)}$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} p(d \mid c)\, p(c)$$

Document Classification

c is the class, d is the document we like to classify

$$c_{MAP} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} p(d \mid c) p(c)$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} p(x_1, x_2, \ldots, x_n \mid c) p(c)$$

Document Classification

c is the class, d is the document we like to classify

$$c_{MAP} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(d \mid c) p(c)$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(x_1, x_2, \ldots, x_n \mid c) p(c)$$

$$\approx \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(x_1 \mid c) p(x_2 \mid c) \ldots p(x_n \mid c) p(c)$$

Conditional Independence Assumption (Naïve Bayes' Rule)

Document Classification

c is the class, d is the document we like to classify

$$c_{MAP} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(d \mid c)\, p(c)$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(x_1, x_2, \ldots, x_n \mid c)\, p(c)$$

$$\approx \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(x_1 \mid c) \times p(x_2 \mid c) \times \ldots \times p(x_n \mid c)\, p(c)$$

And we can easily learn every term on the right hand side from the training data.

Document Classification
Naïve Bayes

c is the class, d is the document we like to classify, x_is are the words.

$$c_{NB} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} p(c) \prod_i p(x_i \mid c)$$

And we can easily learn every term on the right hand side from the training data.

Document Classification, Naïve Bayes: Learning from Training Data

$$c_{NB} = \operatorname*{argmax}_{c \in \{Positive, Negative\}} p(c) \prod_i p(x_i \mid c)$$

$N_{reviews}$: number of reviews
In training set

$$\hat{p}(c_p) = \frac{reviewcount(Positive)}{N_{reviews}}$$

reviewcount(Positive): number of positive reviews.

$$\hat{p}(c_N) = \frac{reviewcount(Negative)}{N_{reviews}}$$

reviewcount(Negative): number of negative reviews.

Document Classification, Naïve Bayes: Learning from Training Data

$$\hat{p}(x_i \mid c_p) = \frac{count(x_i, \textit{positive reviews})}{\sum_{x \in V} count(x, \textit{positive reviews})}$$

Fraction of times word x_i shows up among all words in the positive reviews.

count(x_i, positive reviews): how many times the word x_i has appeared in positive reviews. (you have to repeat this process for every word x_i).

V: The vocabulary. All the words that show up in training reviews.

Document Classification, Naïve Bayes: Learning from Training Data

$$\hat{p}(x_i \mid c_N) = \frac{count(x_i, negative\ reviews)}{\sum_{x \in V} count(x, negative\ reviews)}$$

Fraction of times word x_i shows up among all words in the negative reviews.

count(x_i, negative reviews): how many times the word x_i has appeared in negative reviews. (you have to repeat this process for every word x_i).

V: The vocabulary. All the words that show up in reviews.

Document Classification, Naïve Bayes: Learning from Training Data

Two numerical Challenge:

1- What if a word in a test review has never showed up in positive or negative training reviews?

$$\hat{p}(x_i \mid c_N) = \frac{count(x_i, negative\ reviews)}{\sum_{x \in V} count(x, negative\ reviews)} = 0$$

$$c_{NB} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(c) \prod_i p(x_i \mid c)$$

Document Classification, Naïve Bayes: Learning from Training Data

Two numerical Challenge:

1- What if a word in a test review has never showed up in positive or negative training reviews?

$$\hat{p}(x_i \mid c_N) = \frac{count(x_i, negative\ reviews) + 1}{\sum_{x \in V} (count(x, negative\ reviews) + 1)}$$

$$\hat{p}(x_i \mid c_p) = \frac{count(x_i, positive\ reviews) + 1}{\sum_{x \in V} (count(x, positive\ reviews) + 1)}$$

Document Classification, Naïve Bayes: Learning from Training Data

Two numerical Challenge:

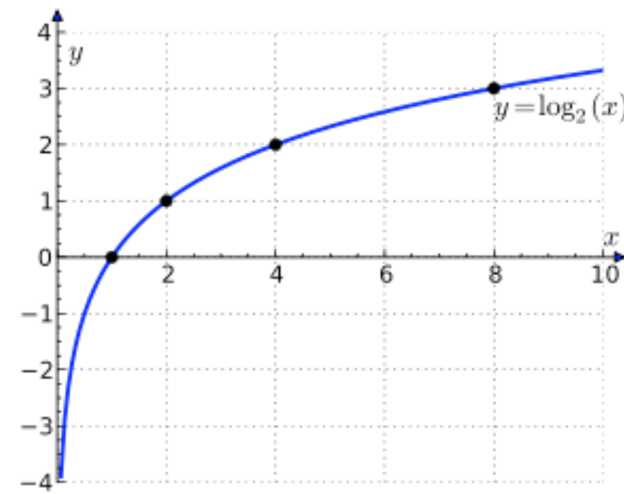
1- What if a word in a test review has never showed up in positive or negative training reviews?
2- We are multiplying many small numbers together. How to fight against underflow?

$$c_{NB} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \; p(c) \prod_i p(x_i \mid c)$$

Document Classification, Naïve Bayes: Learning from Training Data

Two numerical Challenge:

1- What if a word in a test review has never showed up in positive or negative training reviews?
2- We are multiplying many small number fight against underflow?



$$c_{NB} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \log(p(c)\prod_i p(x_i \mid c))$$

Document Classification, Naïve Bayes: Learning from Training Data

Two numerical Challenge:

1- What if a word in a test review has never showed up in positive or negative training reviews?
2- We are multiplying many small numbers together. How to fight against underflow?

$$c_{NB} = \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \log\left(p(c) \prod_i p(x_i \mid c)\right)$$

$$= \underset{c \in \{Positive, Negative\}}{\operatorname{argmax}} \left[\log p(c) + \sum_i \log(p(x_i \mid c))\right]$$

Reading Assigment:

Shimodaira, Hiroshi. "Text classification using naive bayes." *Learning and Data Note* 7 (2014): 1-9.

https://web.stanford.edu/class/cs124/lec/naivebayes.pdf