Bayes and Naïve Bayes Classifier

Aarti Singh

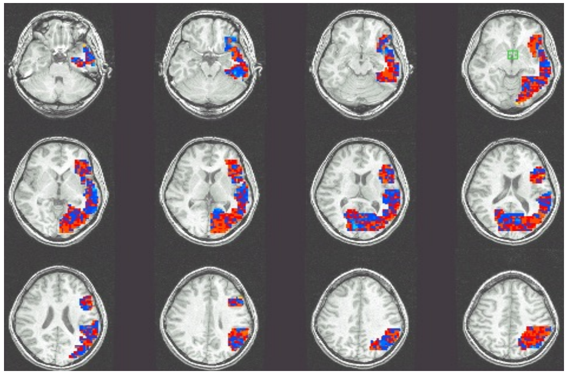
Machine Learning 10-701
Jan 23, 2023

MACHINE LEARNING DEPARTMENT

Carnegie Mellon
School of Computer Science

Classification

<u>Goal</u>: Construct **prediction rule** $f \ : \mathcal{X} \to \mathcal{Y}$



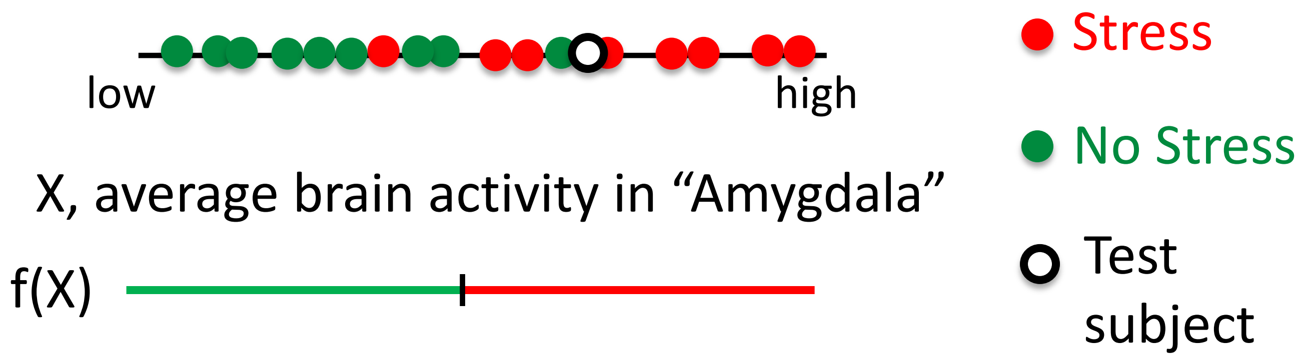
High Stress
Moderate Stress
Low Stress

Input feature vector, X **Label, Y**

In general: label Y can belong to more than two classes
X is multi-dimensional (many features represent an input)

But lets start with a simple case:
label Y is binary (either "Stress" or "No Stress")
X is average brain activity in the "Amygdala"

Binary Classification



Model X and Y as random variables with joint distribution P_{XY}

Training data $\{X_i, Y_i\}^n_{i=1}$ ~ iid (<u>independent</u> and <u>identically distributed</u>) samples from P_{XY}

Test data $\{X,Y\}$ ~ iid sample from P_{XY}

Training and test data are independent draws from **<u>same</u>** distribution

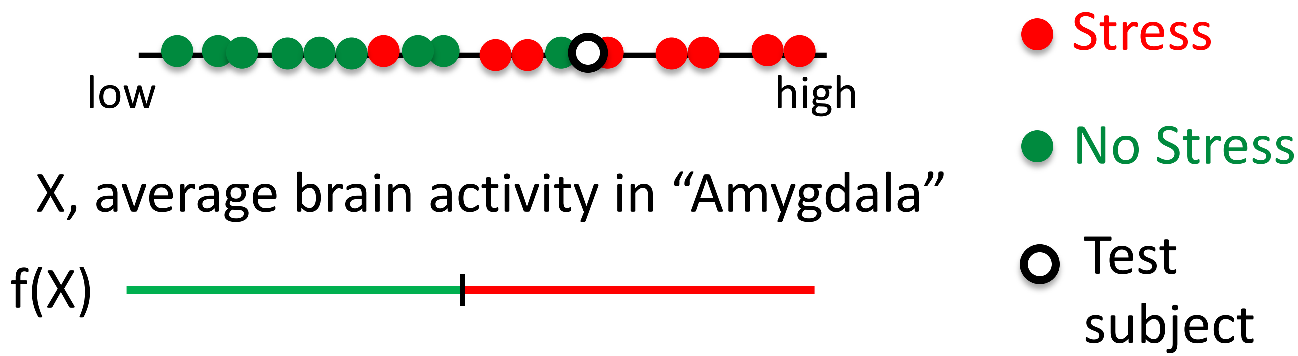
Optimal classifier

Minimize loss in expectation (over random test data)

$$\min_f E_{XY}[\text{loss}(f(X),Y)]$$

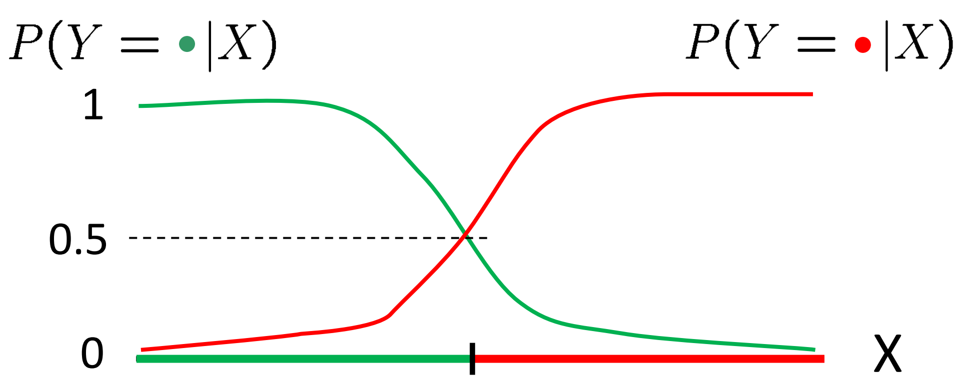
- Which classifier f is optimal for 0/1 loss, assuming we know data-generating distribution P(X,Y)?

Optimal Classifier



X, average brain activity in “Amygdala”

Model X and Y as random variables



For a given X, f(X) = label Y which is more likely

$$f(X) = \arg\max_y P(Y=y|X=x)$$

Optimal classifier

Minimize loss in expectation (over random test data)

$$\min_f E_{XY}[\text{loss}(f(X),Y)]$$

- Which classifier f is optimal for 0/1 loss, assuming we know data-generating distribution P(X,Y)?

Bayes Rule

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

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

To see this, recall:

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

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

Thomas Bayes

Bayes Optimal Classifier

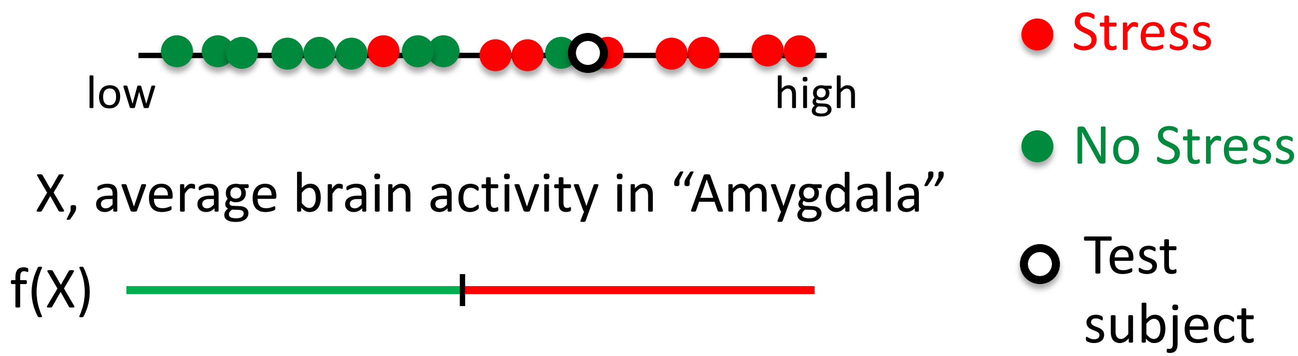
Bayes Rule: $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

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

Bayes optimal classifier:

$$\begin{aligned} \text{f(X)} = \; & \arg\max_{Y=y} P(Y = y|X = x) \\ = \; & \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y = y)}_{\text{Distribution of class}} \end{aligned}$$

Bayes Classifier



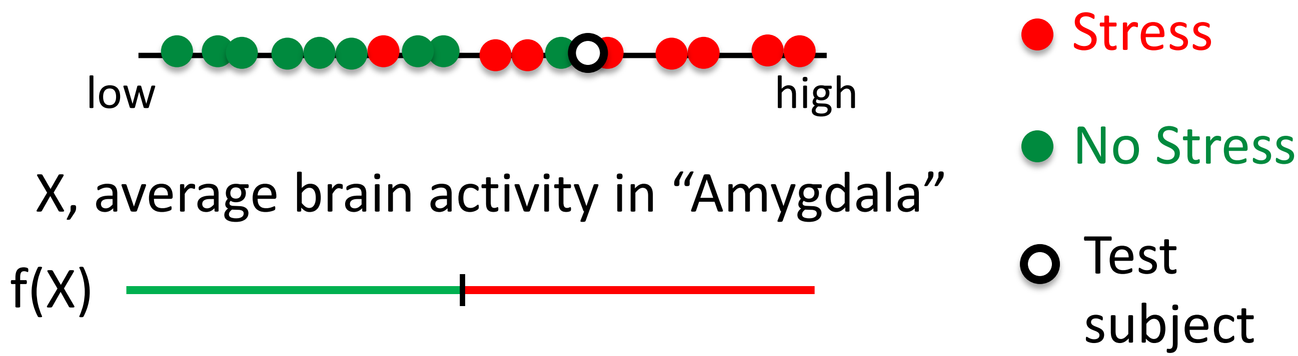
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

We can now consider distribution models to approximate ground truth:

Class distribution P(Y=y)

Class conditional distribution of features P(X=x|Y=y)

Modeling class distribution



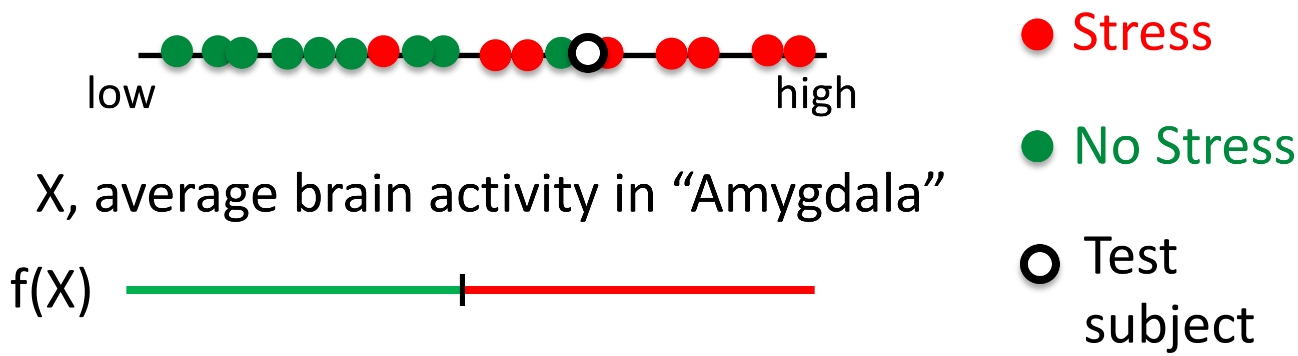
Modeling Class distribution P(Y=y) = Bernoulli(θ)

$$P(Y = {\color{red}\bullet}) = \theta \qquad P(Y = {\color{green}\bullet}) = 1 - \theta$$

Like a coin flip

➢ How do we model multiple (>2) classes?

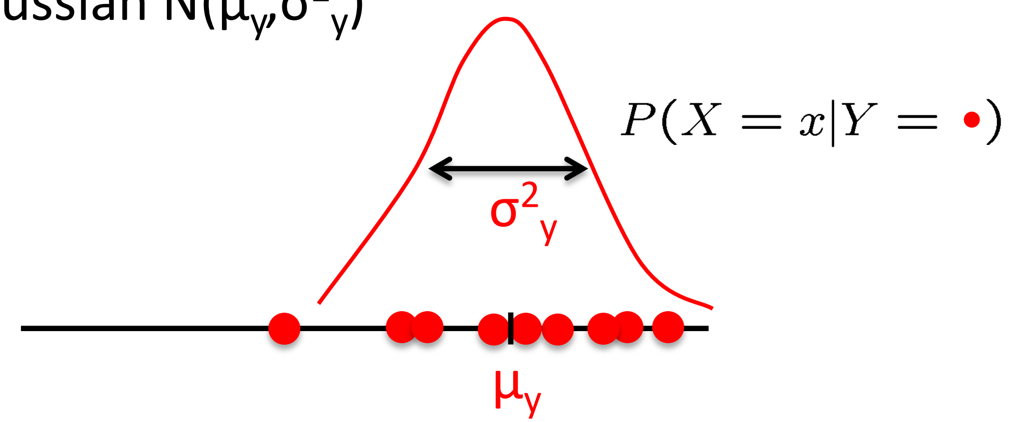
Modeling class conditional distribution of features



Modeling class conditional distribution of feature P(X=x|Y=y)

➢ What distribution would you use?

E.g. P(X=x|Y=y) = Gaussian $N(\mu_y, \sigma^2_y)$



Gaussian Bayes classifier

$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

Class conditional
Distribution of features

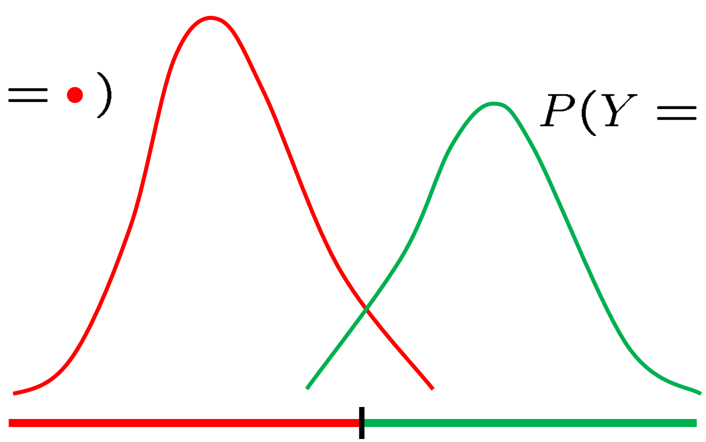
Class distribution

Use MLE/MAP to learn parameters θ, μ_y, Σ_y from data

Gaussian(μ_y, Σ_y)

Bernoulli(θ)

$P(Y=\bullet)P(X=x|Y=\bullet)$



$P(Y=\bullet)P(X=x|Y=\bullet)$

Poll

- Is the Gaussian Bayes Classifier always optimal under 0/1 loss?

  A. True B. False

1-dim Gaussian Bayes classifier

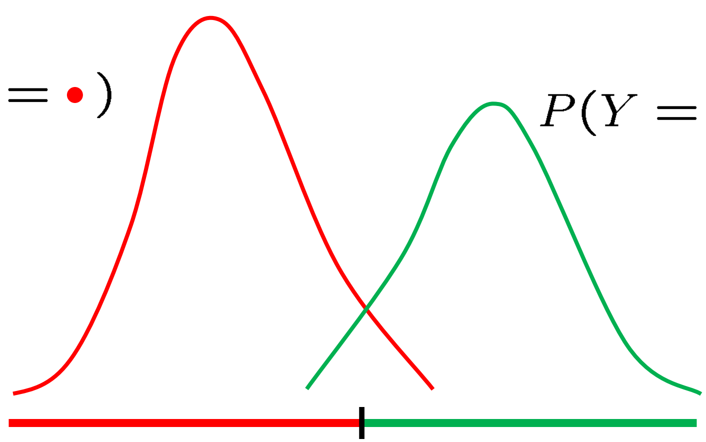
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

Gaussian(μ_y, σ^2_y)

Bernoulli(θ)

- What decision boundaries can we get in 1-dim?

$P(Y=\bullet)P(X=x|Y=\bullet)$



$P(Y=\bullet)P(X=x|Y=\bullet)$

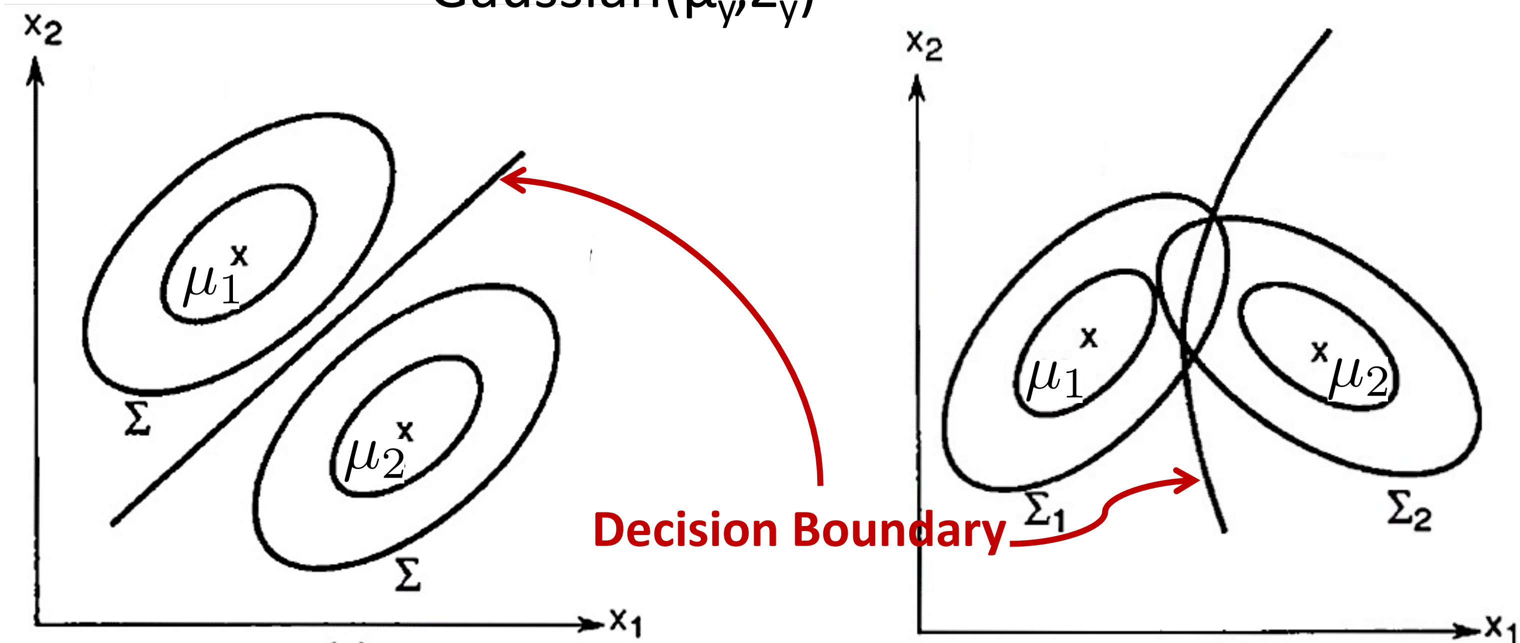
d-dim Gaussian Bayes classifier

$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

- What decision boundaries can we get in d-dim?

Class conditional Distribution of features → Gaussian(μ_y,Σ_y)

Class distribution → Bernoulli(θ)



Decision Boundary of Gaussian Bayes

- Decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

Compute the ratio

$$1 = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}$$

In general, this implies a quadratic equation in x. But if $\Sigma_1 = \Sigma_0$, then quadratic part cancels out and decision boundary is linear.

Recap

- **Bayes classifier** – assumes P_{XY} known, optimal for 0/1 loss

$$f(X) = \arg\max_{Y=y} P(Y=y|X=x)$$

$$= \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

- **Gaussian Bayes classifier** – assumes
  Class distribution is Bernoulli/Multinomial
  Class conditional distribution of features is Gaussian

- **Decision boundary** – (binary classification)

How many parameters do we need to learn (continuous features)?

Class distribution:

$P(Y = y) = p_y$ for all y in H, M, L p_H, p_M, p_L (sum to 1)

K-1 if K labels

Class conditional distribution of features:

$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y μ_y – d-dim vector

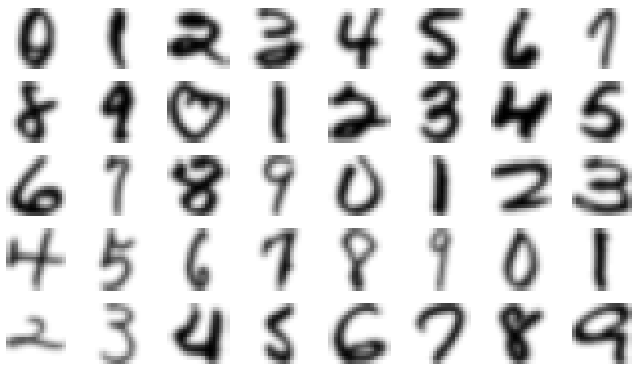
Σ_y - dxd matrix

$Kd + Kd(d+1)/2 = O(Kd^2)$ if d features

Quadratic in dimension d! If d = 256x256 pixels, ~ 13 billion parameters!

How many parameters do we need to learn (discrete features)?

Class distribution:



$P(Y = y) = p_y$ for all y in 0, 1, 2, …, 9

$p_0, p_1, \ldots, p_9$ (sum to 1)

K-1 if K labels

Class conditional distribution of (binary) features:

$P(X=x|Y = y)$ ~ For each label y, maintain probability table with 2^d-1 entries

$K(2^d – 1)$ if d binary features

Exponential in dimension d!

What's wrong with too many parameters?

- How many training data needed to learn one parameter (bias of a coin)?

- Need lots of training data to learn the parameters!
  - Training data > number of (independent) parameters

Naïve Bayes Classifier

- Bayes Classifier with additional "naïve" assumption:
  - Features are independent given class:

$$\begin{aligned} P(X^{(1)}, X^{(2)}|Y) &= P(X^{(1)}|X^{(2)}, Y)P(X^{(2)}|Y) \\ &= P(X^{(1)}|Y)P(X^{(2)}|Y) \end{aligned}$$

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$$

  - More generally:

$$P(X^{(1)}, ..., X^{(d)}|Y) = \prod_{i=1}^{d} P(X^{(i)}|Y)$$

$$X = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(d)} \end{bmatrix}$$

- If conditional independence assumption holds, NB is optimal classifier! But worse otherwise.

Conditional Independence

- X is **conditionally independent** of Y given Z:

  probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

- e.g., $P(Thunder | Rain, Lightning) = P(Thunder | Lightning)$

  Note: does NOT mean Thunder is independent of Rain

Naïve Bayes Classifier

- Bayes Classifier with additional "naïve" assumption:
  - Features are independent given class:

$$P(X^{(1)}, \ldots, X^{(d)}|Y) \;=\; \prod_{i=1}^{d} P(X^{(i)}|Y)$$

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_{y} \; P(x^{(1)}, \ldots, x^{(d)}|y)\; P(y) \\ &= \arg\max_{y} \; \prod_{i=1}^{d} P(x^{(i)}|y)\, P(y) \end{aligned}$$

- How many parameters now?

How many parameters do we need to learn (continuous features)?

- Poll

Number of parameters for class distribution P(Y=y) for K classes?

Number of parameters for Class conditional distribution of features P(X = x|Y = y) for d features (using Gaussian Naïve Bayes assumption)?

A. K-1, Kd

B. K-1, K(d + d(d+1)/2)

C. K-1, Kd^2

D. K-1, 2Kd

How many parameters do we need to learn (discrete features)?

- Poll

Number of parameters for class distribution P(Y=y) for K classes?

Number of parameters for Class conditional distribution of features P(X = x|Y = y) for d binary features (using Naïve Bayes assumption)?

A. K-1, $K2^d$

B. K-1, K(d-1)

C. K-1, Kd

D. K-1, 2Kd

Naïve Bayes Classifier

- Bayes Classifier with additional "naïve" assumption:
  - Features are independent given class:

$$P(X^{(1)}, ..., X^{(d)}|Y) = \prod_{i=1}^{d} P(X^{(i)}|Y)$$

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x^{(1)}, ..., x^{(d)}|y) \; P(y) \\ &= \arg\max_y \; \prod_{i=1}^{d} P(x^{(i)}|y) \, P(y) \end{aligned}$$

- Has fewer parameters, and hence requires fewer training data, even though assumption may be violated in practice

Learned Gaussian Naïve Bayes Model

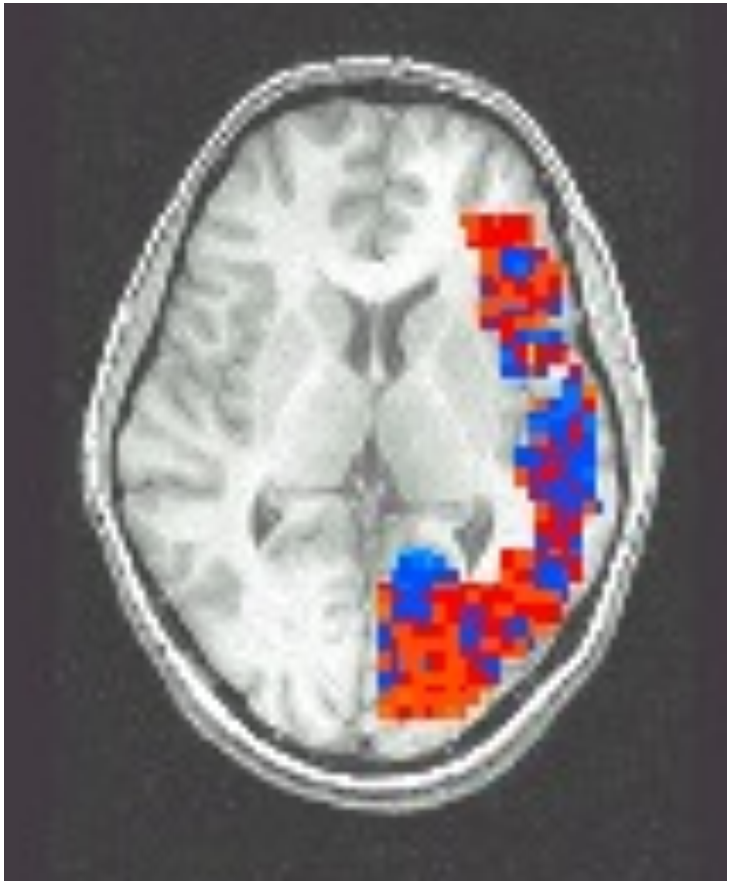
Means for P(BrainActivity | WordCategory)

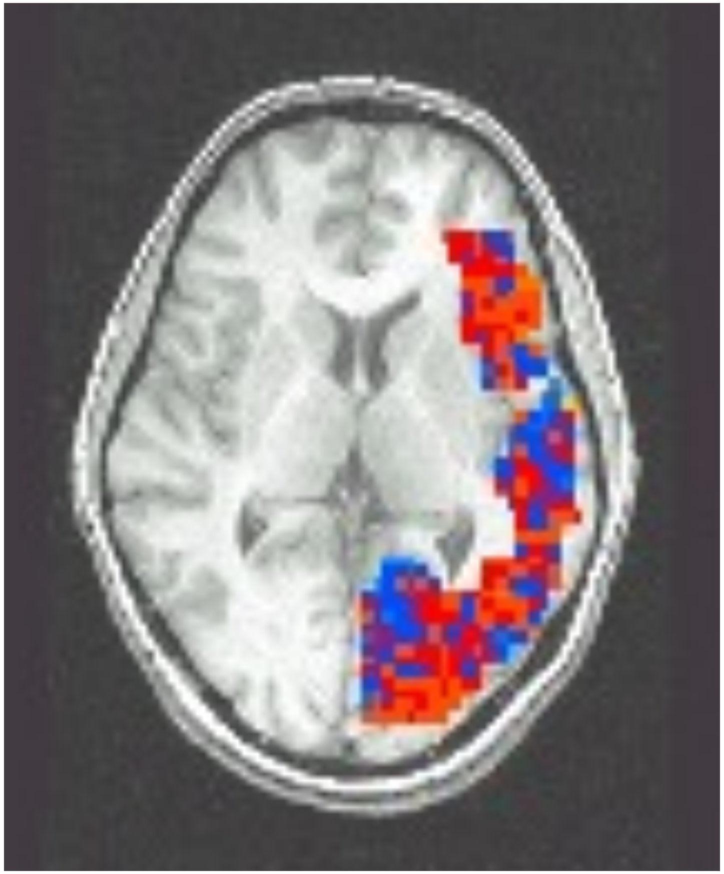
[Mitchell et al.03]

Pairwise classification accuracy: 85%

People words

Animal words





Text classification

Input $X \in \mathcal{X}$

remember to wake up when class ends
=
wake ends to class remember up when

How to represent inputs mathematically?

- Document vector X
  - list of words (different length for each document)
  - frequency of words (length of each document = size of vocabulary), also known as **Bag-of-words** approach

    Misses out context!!
  - list of n-grams (n-tuples of words)

➢ Ideas?

➢ Why might this be limited?

Text classification

Raw input ⇨ Features ⇨ Model for input features


Third Edition April 28, 2020

The Quarantine Times

Kids Connecting to their Community

Dear Future Me, Letters to remind us all what life under quarantine was like

Dear future self,
Right now we are Quarantined because of Covid-19, and we can not go anywhere! I have been reading and riding my bike so much lately! Today is Earth day, so I am going to spend lots of time outside!
Nani
3rd Grade, Edgewood Campus School

Dear 20 year old Claire,
Hi! What do you remember about Covid-19? Today I got a video that showed the 4th grade classroom. I was happy to see my class. I got to talk to my friend. She was biking across the street!
My mom and I started a newspaper called "Quarantine Times." I miss all my friends. I have a few questions for you (me). What's your job? How is it like being 20?

word1	5
word2	2
word3	10
word4	20
word5	12
word6	5
word7	8
word8	4
.	.
.	.
.	.

P(X=x|Y=y)
= P(word1 = 5, word2 = 2, word3 = 10, … | Y=y)

Bayes classifier: $\arg\max_y \; P(x^{(1)}, \ldots, x^{(d)} | y) \; P(y)$

Naïve Bayes classifier: $\arg\max_y \; \prod_{i=1}^{d} P(x^{(i)} | y) \, P(y)$

Glossary of Machine Learning

- iid random variables
- Class prior
- Class conditional distribution of inputs
- Optimal classifier under 0/1 loss
- Bayes rule
- Gaussian Bayes classifier
- Naïve Bayes classifier
- Decision boundary