Learning Distributions
Maximum Likelihood Estimate (MLE)
Bayes Classifier

Aarti Singh

Machine Learning 10-315
Sept 9, 2020

Logistics

- Anonymous feedback form
- Recitation on Friday Sept 11 – MLE/MAP + Optimization methods review and hands-on exercises
- QnA1 due TODAY
- HW1 to be released TODAY

Why is ML not …

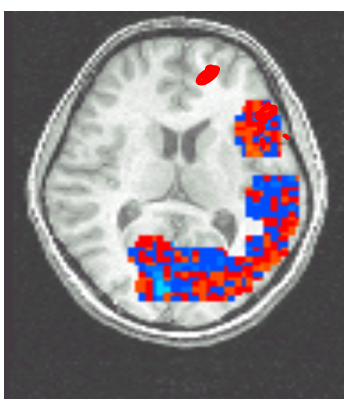
- Interpolation?
  - Noise, stochasticity, transfer across domains, …
- Statistics?
  - care about computationally efficiency (feasible, at least polynomial time in input size but typically much faster)
- Optimization?

  $\mathbb{E}_{XY}[\text{loss}(f(x), y)]$
  - Don't know true objective function, only stochastic version computed using data samples
- Data mining?
  - Generalization on new unseen data
- Your question?

Unsupervised Learning

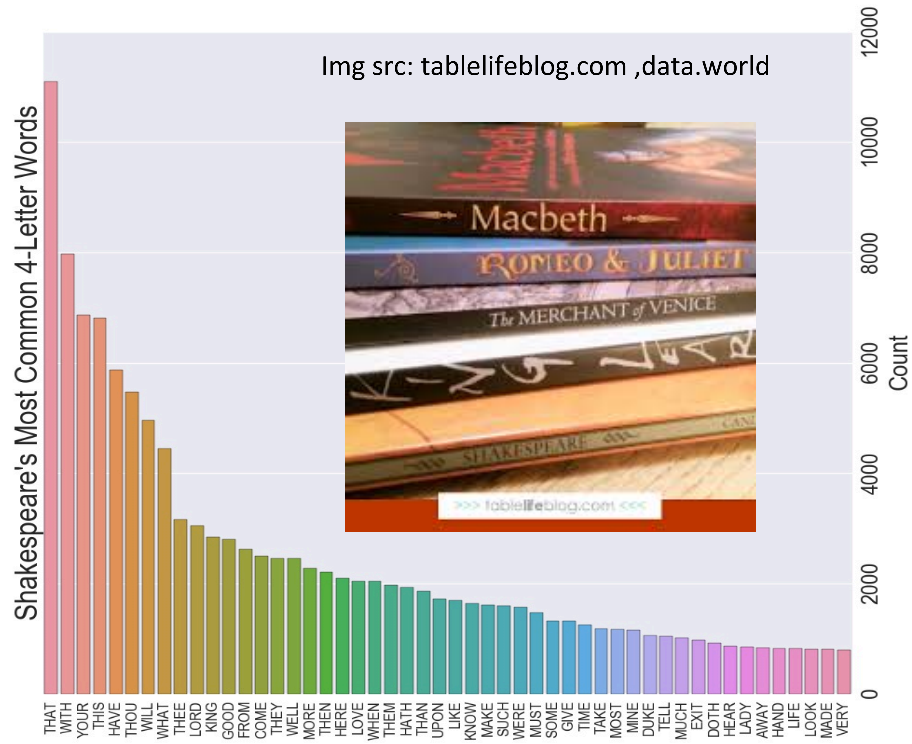
Learning a Distribution

Bias of a coin



voxel

Distribution of brain activity under stimuli



Distribution of words in text

Notion of "Features aka Attributes"

Input $X \in \mathcal{X}$

remember to wake up when class ends

=

wake ends to class remember up when

How to represent inputs mathematically?

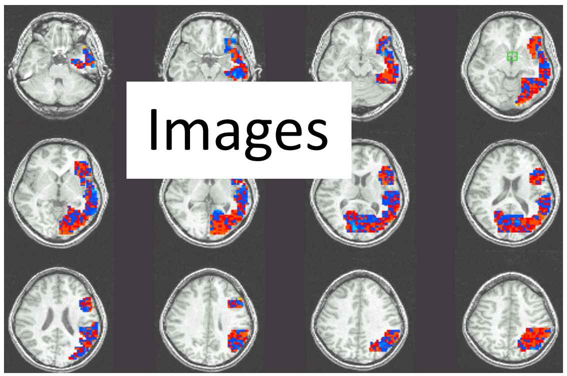
$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$

- Document vector X
  - ➢ Ideas?
  - 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)
    - $n=2$ $d=|Voc|^2$
    - $d\uparrow$ as $n\uparrow$

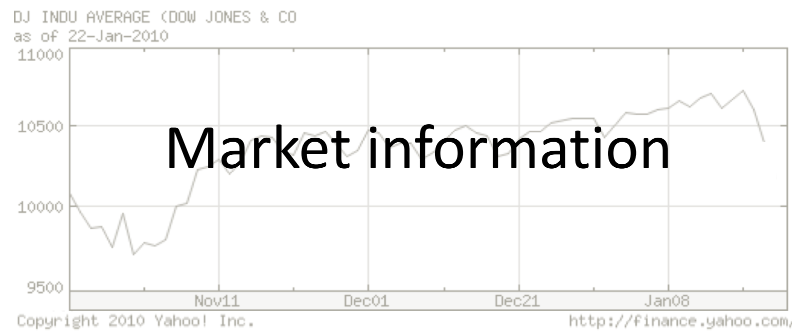
➢ Why might this be limited?

Notion of "Features aka Attributes"

Input $X \in \mathcal{X}$



Input $X \in \mathcal{X}$



$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ $x_i \in \mathbb{R}$

How to represent inputs mathematically?

- Image X = intensity/value at each pixel, fourier transform values, SIFT etc.
- Market information X = daily/monthly? price of share for past 10 years

Distribution of Inputs

Input $X \in \mathcal{X}$

Discrete Probability Distribution P(X) = P(X=x)

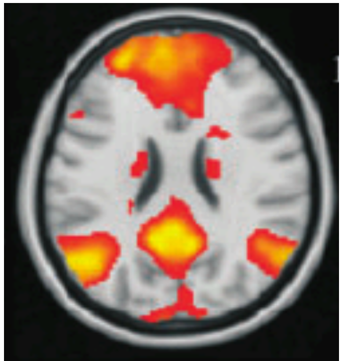
e.g. P(head) = ½, P(word x in text) = p_x

Probabilities in a distribution sum to 1

$p(X=x) \geq 0$ $\sum_x P(X=x) = 1$ P(tail) = 1 – p(head), $\sum_x p_x = 1$

Continuous Probability density p(x) P(a<=X<=b) = $\int_a^b p(x)dx$

e.g. p(brain activity)



Probability density integrate to 1

$p(x) \geq 0$ & $\int p(x)dx = 1$

Distributions in Supervised tasks

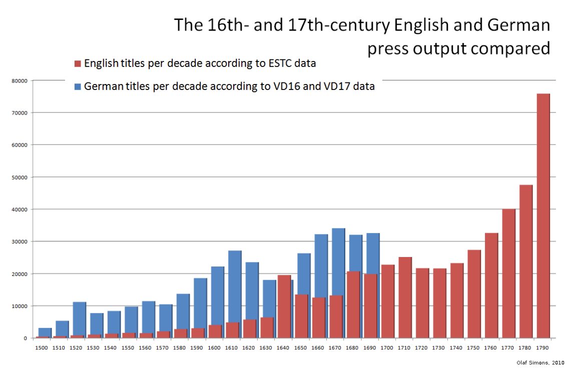
Input $X \in \mathcal{X}$

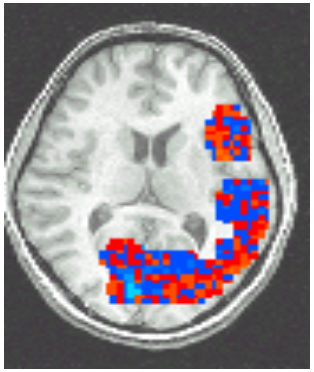
- Distribution learning also arises in supervised learning tasks e.g. classification

P(Y= y) topic — Distribution of class labels

P(X = x | Y = y) — Distribution of words in 'news' documents

'news' — Distribution of brain activity under 'stress'



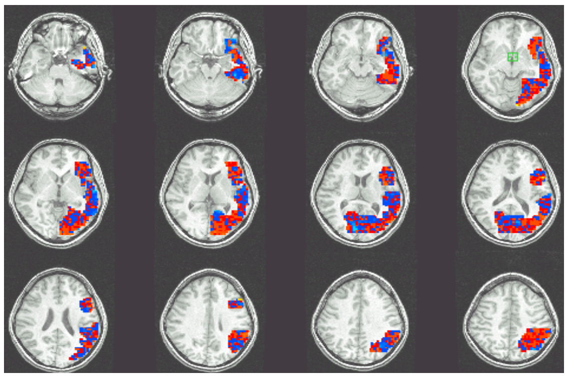


Olaf simons'10

P(Y = y|X = x) — Distribution of topics given document

Classification

Goal: Construct **prediction rule** $f \; : \mathcal{X} \to \mathcal{Y}$



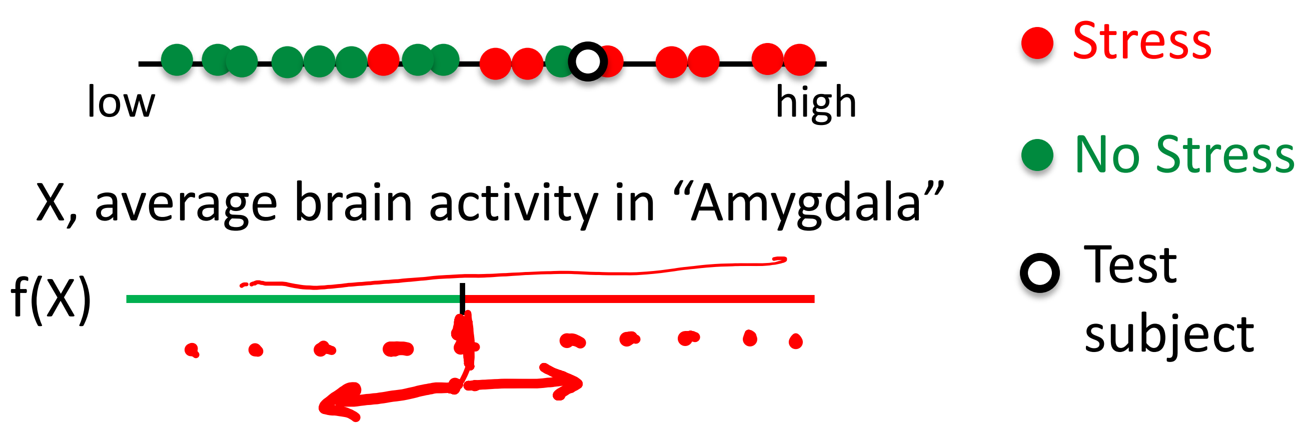
High Stress
Moderate Stress
Low Stress

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” $= X \in \mathbb{R}$

Binary Classification



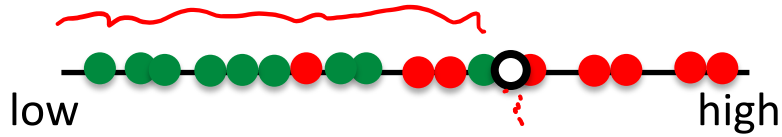
Model X and Y as random variables with joint distribution P_{XY} $= p(X, Y)$

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

Bayes Classifier

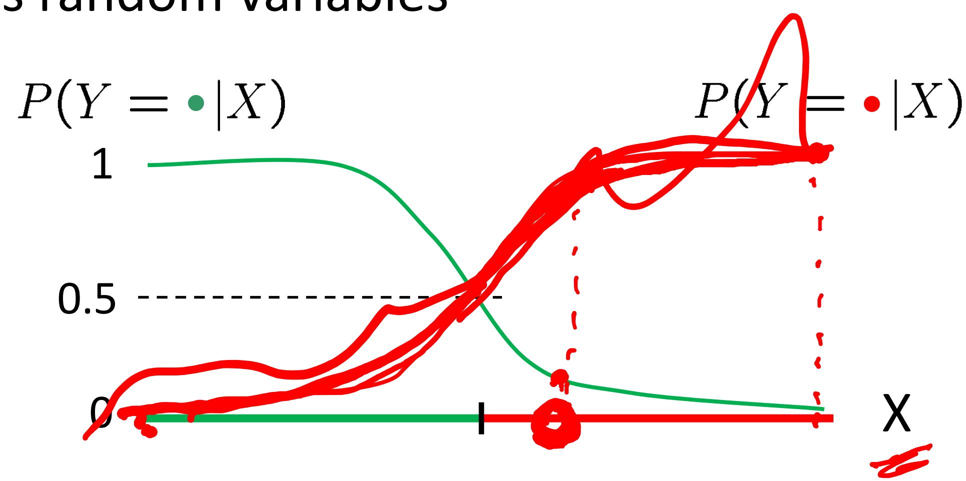


X, average brain activity in “Amygdala”

- Stress
- No Stress
- Test subject

f(X) 

Model X and Y as random variables



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

P(Y|X)

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

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)

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

Thomas Bayes

Bayes Classifier

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

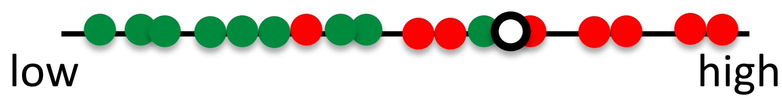
$$P(Y = y|X = x) \;=\; \underbrace{\frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}}$$

Bayes classifier:

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

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

Bayes Classifier



Stress

No Stress

Test subject

X, average brain activity in “Amygdala”

Optimal

$E[\text{loss}(f(x), y)] \rightarrow$

loss = 0/1

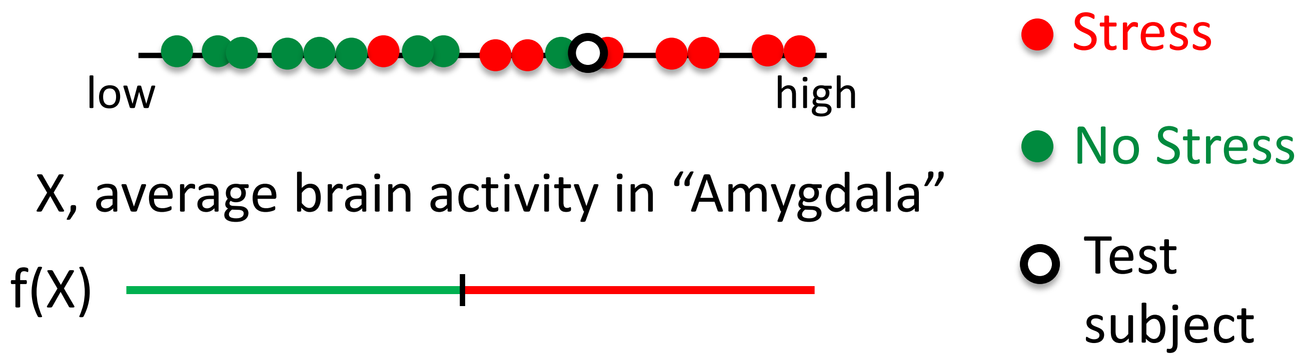
$$f(X) = \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y = y)}_{\text{Class distribution}}$$

We can now consider appropriate distribution models for the two terms:

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)

$$P(Y = \bullet) = \theta \quad \text{parameter} \qquad P(Y = \bullet) = 1 - \theta$$

Like a coin flip

How to learn parameters from data?
MLE

(Discrete case)

Learning parameters in distributions

$$P(Y = \bullet) = \theta \qquad\qquad P(Y = \bullet) = 1 - \theta$$

Learning θ is equivalent to learning probability of head in coin flip.

- How do you learn that?

Data = 

Answer: 3/5 ← frequency of heads

- Why??

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data aka Likelihood</u>

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \overbrace{P(D \mid \theta)}^{J(\theta)}$$

D ~ iid draw from Ber(θ)

MLE of probability of head:

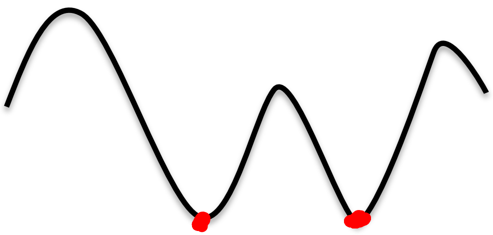
count of H

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5**

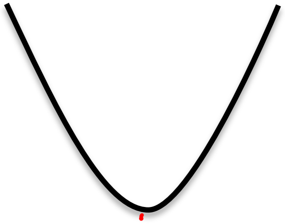
"Frequency of heads"

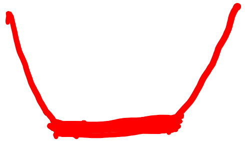
Short detour - Optimization

- Optimization objective $J(\theta)$
- Minimum value $J^* = \min_\theta J(\theta)$ or $\max_\theta J(\theta)$
- Minima (points at which minimum value is achieved) may not be unique

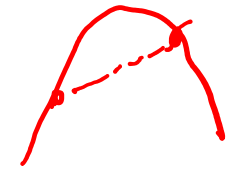


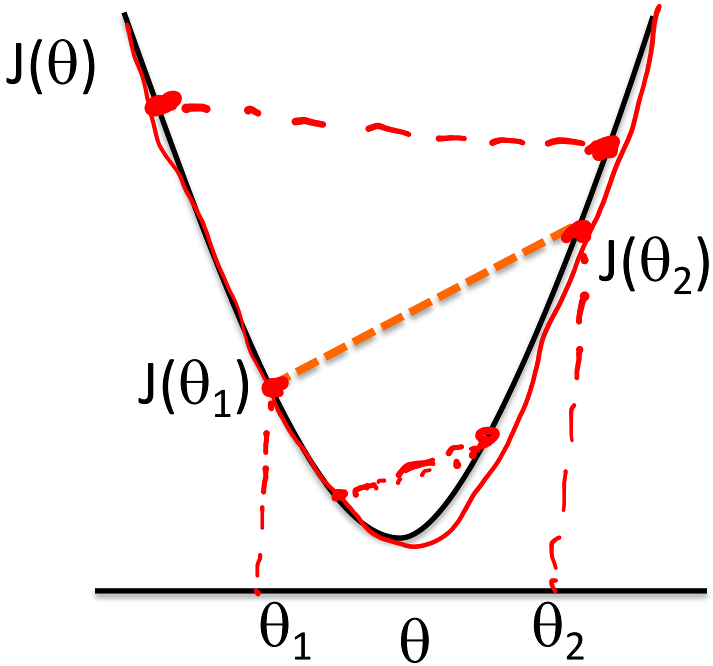
- If function is strictly convex, then minimum is unique





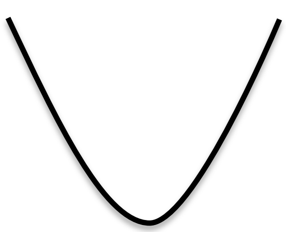
Convex functions



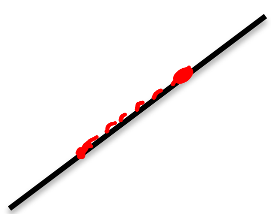


A function $J(\theta)$ is called **convex** if the line joining two points $J(\theta_1), J(\theta_2)$ on the function does not go below the function on the interval $[\theta_1, \theta_2]$

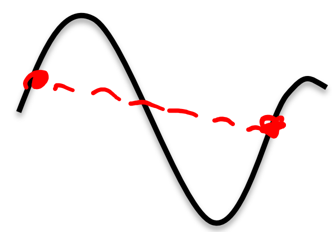
(Strictly) Convex functions have a unique minimum!



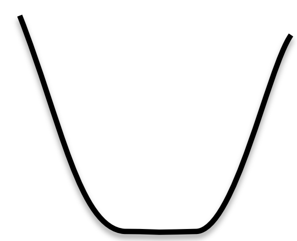
Convex



Both Concave & Convex



Neither

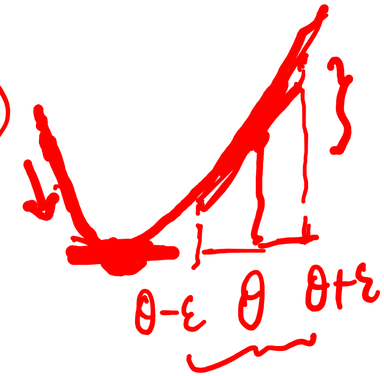


Convex but not strictly convex

Optimizing convex (concave) functions

- Derivative of a function

$$\frac{\Delta J(\theta)}{\Delta \theta} = \lim_{\epsilon \to 0} \frac{J(\theta+\epsilon) - J(\theta-\epsilon)}{2\epsilon}$$



  - Partial derivative $\frac{\partial J(\theta)}{\partial \theta}$

- Derivative is zero at minimum of a convex function

$$g(\theta) = \left. \frac{\partial J(\theta)}{\partial \theta} \right|_{\theta = \theta^*} = 0$$

- Second derivative is positive at minimum of a convex function

$$\frac{\partial g(\theta)}{\partial \theta} = \frac{\partial^2 J(\theta)}{\partial \theta^2} \geq 0$$

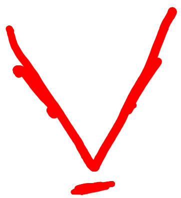
Optimizing convex (concave) functions

- What about

  concave functions?

  non-convex/non-concave functions?

  functions that are not differentiable?

  optimizing a function over a bounded domain aka constrained optimization?

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \underline{P(D \mid \theta)}$$

$J(\theta)$

① D ~ iid

② Ber(θ)

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

= 3/5

"Frequency of heads"