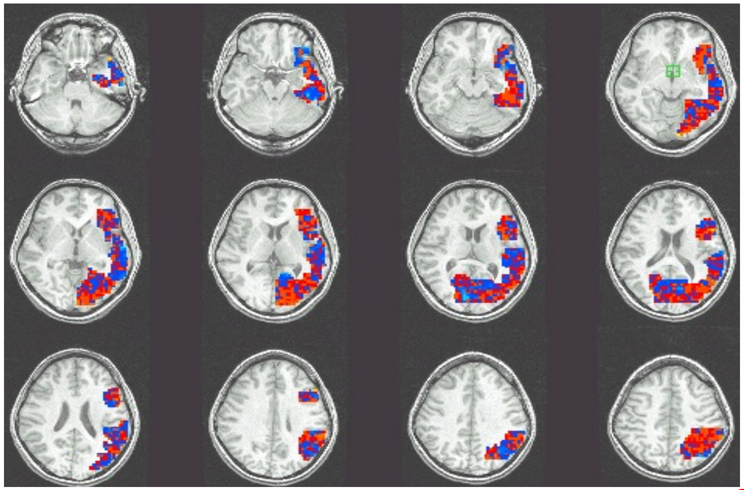
Naïve Bayes contd…

Aarti Singh

Machine Learning 10-315
Feb 2, 2022

Multi-class, multi-dimensional classification – Continuous features



High Stress
Moderate Stress
Low Stress

Input feature vector, X = $\begin{bmatrix} x_{(1)} \\ x_{(d)} \end{bmatrix}$ **Label, Y**

We started with a simple case:

label Y is binary (either "Stress" or "No Stress")

X is average brain activity in the "Amygdala"

In general: label Y can belong to K>2 classes

X is multi-dimensional d>1 (average activity in all brain regions)

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

Class probability:

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

K-1 if K labels

Class conditional distribution of features: $\qquad Kd + K\frac{d(d+1)}{2}$

$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y $\qquad$ d×d $\qquad$ y

μ_y – d-dim vector

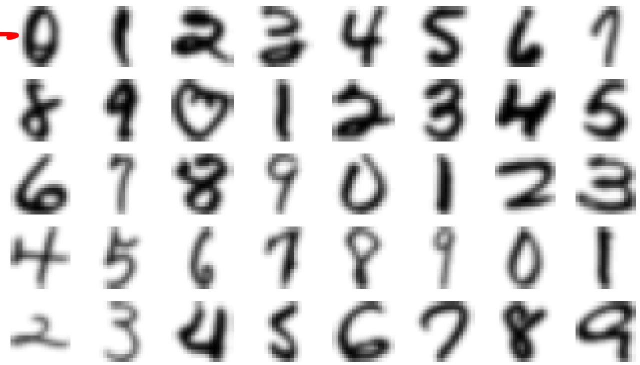
Σ_y - dxd matrix

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

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

Multi-class, multi-dimensional classification - Discrete features

$$X = \begin{array}{cc} \begin{bmatrix} 0|1 \\ 0|1 \\ \vdots \\ 0|1 \end{bmatrix} & \begin{array}{c} X_{(1)} \\ \vdots \\ \\ X_{(d)} \end{array} \end{array}$$



“0”
“1”
…
“9”

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

$$X = \begin{array}{cc} \begin{bmatrix} - \\ - \\ - \\ - \\ - \end{bmatrix} & \begin{array}{c} X_{(1)} \\ \vdots \\ \\ X_{(d)} \end{array} \end{array}$$

Sports
Science
News

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

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

Class probability:

$P(Y = y) = p_y$ for all y in 0, 1, 2, …, 9 $\quad$ $p_0, p_1, \ldots, p_9$ (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

$\begin{bmatrix} 0/1 \\ \vdots \\ 0/1 \end{bmatrix} d$

$K(2^d – 1)$ 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:

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

$$P(X|Y) = P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$

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

$$= P(X_1|Y) \cdot P(X_2|Y)$$

  - More generally:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_d \end{bmatrix}$$

$$P(X|Y) = P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

$X = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$

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

Conditional Independence

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

- 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

$\left(X = \begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(d)} \end{bmatrix} \right.$ Y = low stress $\Big)$ $X_1 \ldots X_n$ $\begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(d)} \end{bmatrix}$

Naïve Bayes Classifier

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

$d \times d$

$\Sigma_y = \begin{bmatrix} \sigma_{(1)}^2 & 0 \\ 0 & \sigma_{(d)}^2 \end{bmatrix}$

$\sim N(\mu_y, \Sigma_y)$

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

$\Sigma_y(i,j) = E[(X_i - E X_i)(X_j - E X_j)] \mid Y$

$$f_{NB}(\mathbf{x}) = \arg\max_y P(x_{(1)}, \ldots, x_{(d)} \mid y) P(y)$$

$$= \arg\max_y \prod_{i=1}^{d} P(x_{(i)}|y) P(y)$$

- How many parameters now?

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

$E[(X_i - EX_i) | Y] = 0$

- Poll

$P(X|Y) \sim \mathcal{N}(\mu_y, \Sigma_y)$ $\quad (\Sigma_y)_{ij} = E[(X_i - EX_i)(X_j - EX_j) | Y]$

$$\Sigma_y = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} \quad d \quad \times K$$

$+$

$$\mu_y = \begin{bmatrix} \\ \\ \end{bmatrix} \quad d \quad \times K$$

vs. Kd^2

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

$E[AB] = E[A]E[B]$

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

- Poll

$$P(X_1=x_1, X_2=x_2, \ldots X_d=x_d \mid Y)$$

$$= \prod_{i=1}^{d} \underbrace{P(X_i=x_i \mid Y)}_{d}$$

vs. $O(K2^d)$

Naïve Bayes Classifier

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

$$P(X_1 \ldots X_d | Y) = \prod_{i=1}^{d} P(X_i | Y)$$

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y) P(y) \\ &= \arg\max_y \prod_{i=1}^{d} P(x_i | y) P(y) \end{aligned}$$

- 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)

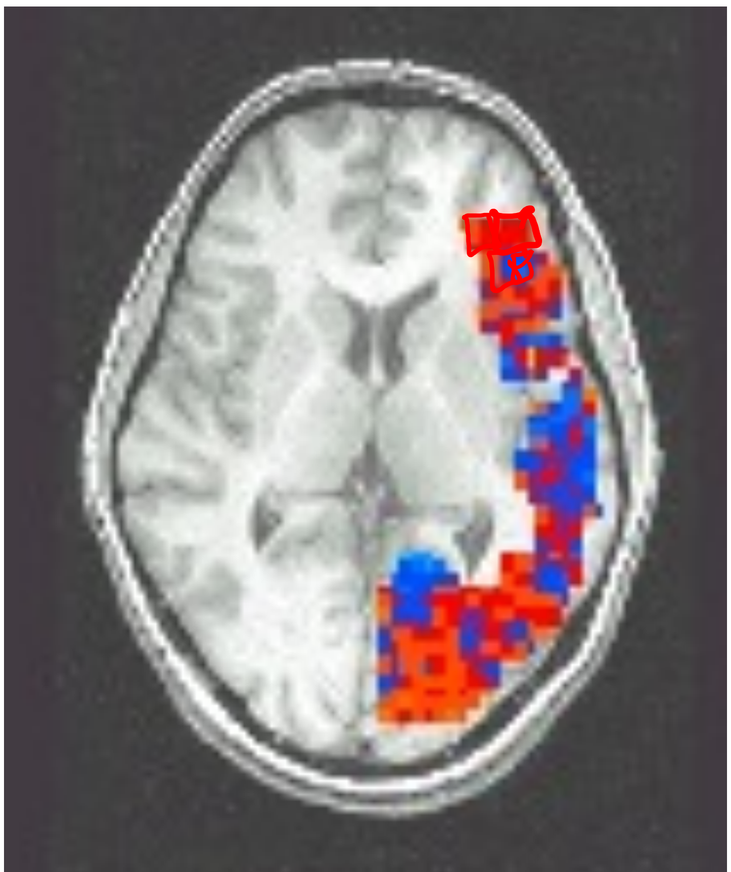
[Mitchell et al.03]

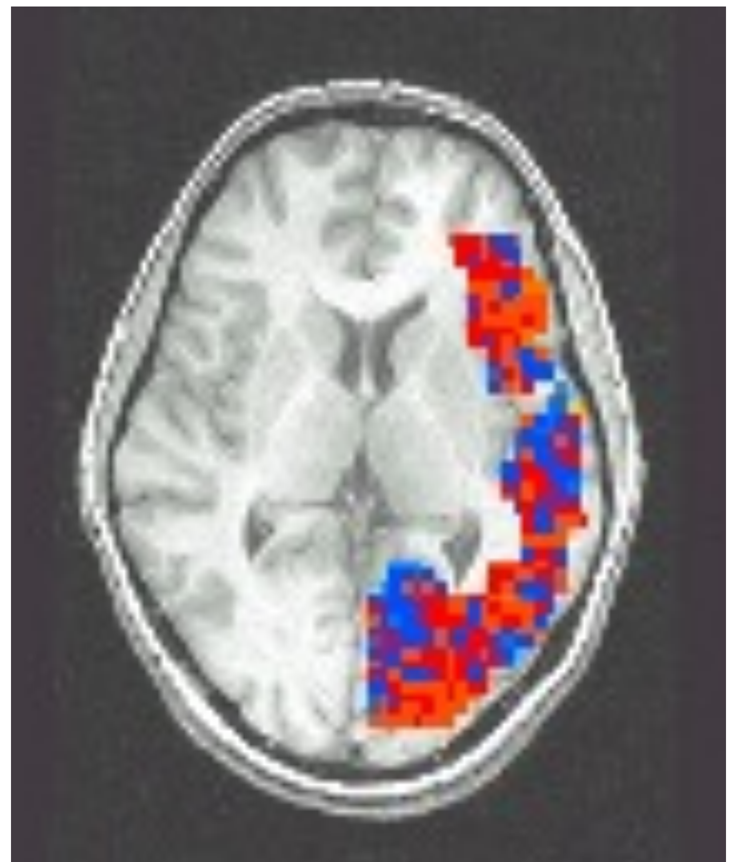
μ_y Σ_y

Pairwise classification accuracy: 85%

People words

Animal words





Text classification

Raw input , X → Features $\begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(d)} \end{bmatrix}$ → 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
.	.
.	.
.	.

Bag of words

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

word $|V|$

$$= \prod_{i=1}^{|V|} P(X_{(i)} = x_{(i)} \mid Y=y)$$

HW1!

Bag of words + Naïve Bayes