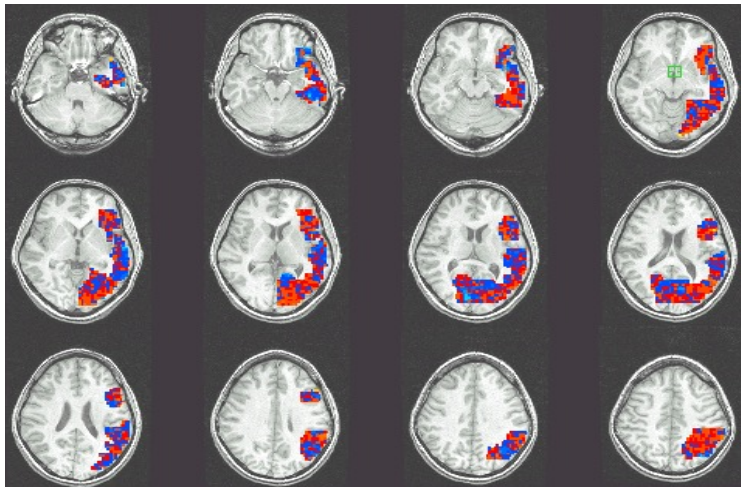
Naïve Bayes

Aarti Singh

Machine Learning 10-315
Sept 13, 2021

Multi-class, multi-dimensional classification – Continuous features



High Stress
Moderate Stress
Low Stress

Input feature vector, X

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 p_H, p_M, p_L (sum to 1)

K-1 if K labels

Class conditional distribution of features:

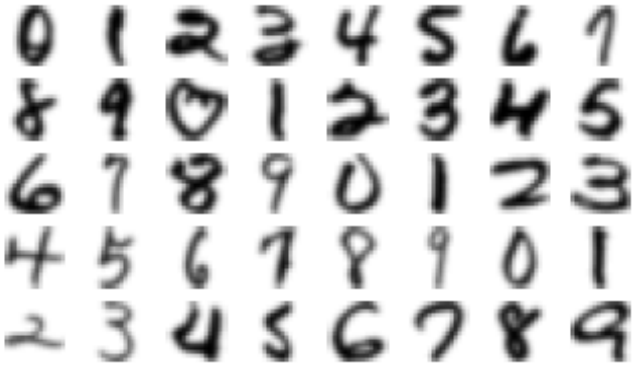
$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y μ_y – d-dim vector

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!

Multi-class, multi-dimensional classification - Discrete features



“0”
“1”
…
“9”

Input feature vector, X

Label, Y

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 $p_0, p_1, \ldots, p_9$ (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:

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

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

  - More generally:

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

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

- 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

Conditional vs. Marginal Independence

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Wearing coats is independent of accidents conditioning on the fact that it rained

Naïve Bayes Classifier

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

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

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

- How many parameters now?

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

- Poll

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

- Poll

Naïve Bayes Classifier

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

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

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

- 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]

Pairwise classification accuracy: 85%

People words

Animal words