

Classification - Naïve Bayes

Aarti Singh

Co-instructor: Barnabas Poczos

Machine Learning 10-401

Jan 19, 2016



MACHINE LEARNING DEPARTMENT



Announcements

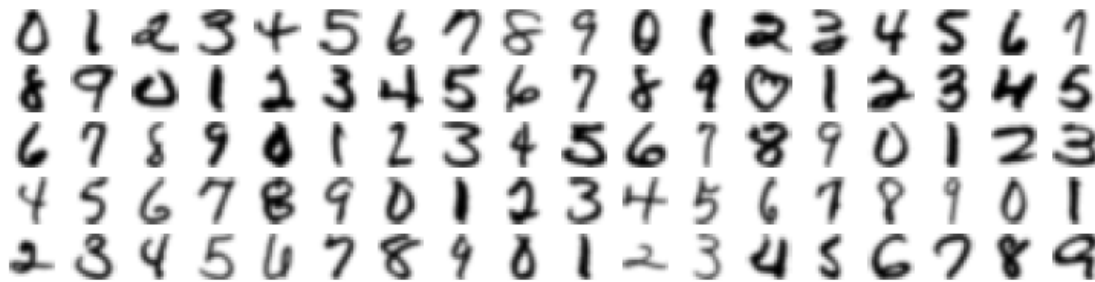
- Recitation on Wed 6:30-7:30 pm GHC 4303
 - By Yining Wang
 - More multi-variate probability and linear algebra (vector/matrix manipulation)
- HW0

Hand-written digit recognition

Input, X (images of hand-written digits)



Label, Y



0, 1, 2, ..., 9

Feature representation:

Grey-scale images – d-dim vector of d pixel intensities

Continuous features

Black-white images – d-dim binary (0/1) vector

Discrete features

Bayes Optimal Classifier

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

Class conditional density Class probability

Model parameters in hand-written digits recognition:

$$P(Y=y) = p_y \text{ for } y = 0, 1, \dots, 9$$

$$p_9 = 1 - (p_0 + p_1 + \dots + p_8)$$

K-1 if K labels

Grey-scale images – d-dim vector of d pixel intensities

$$P(X=x|Y=y) \sim N(\mu_y, \Sigma_y) \text{ for each } y$$

$Kd + Kd(d+1)/2$: Quadratic in dimension d!

Black-white images – d-dim binary (0/1) vector

$$P(X=x|Y=y) \sim \text{probability table with } 2^d - 1 \text{ entries for each } y$$

$K(2^d - 1)$: Exponential in dimension d!

Need lots of training data!!

Naïve Bayes Classifier

- Bayes Classifier with additional “naïve” assumption:

- Features are independent given class:

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

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

- More generally:

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

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

- 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:

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

- e.g., $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

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:

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

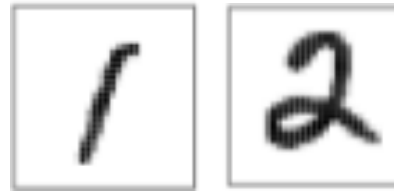
$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg \max_y P(x_1, \dots, x_d | y) P(y) \\ &= \arg \max_y \prod_{i=1}^d P(x_i | y) P(y) \end{aligned}$$

- How many parameters now?

Handwritten digit recognition (continuous features)

Training Data:

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



... n greyscale
images with
d pixels

Y

1

2

... n labels

How many parameters?

Class probability $P(Y = y) = p_y$ for all y **K-1 if K labels**

May not
hold

Class conditional distribution of features (using Naïve Bayes assumption)

$P(X_i = x_i | Y = y) \sim N(\mu_i^{(y)}, \sigma_i^{2(y)})$ for each y and each pixel i **2Kd**

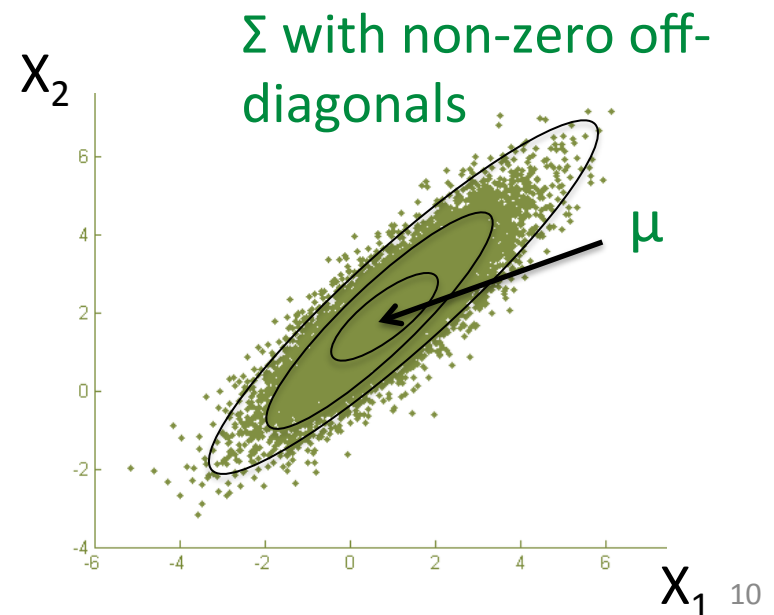
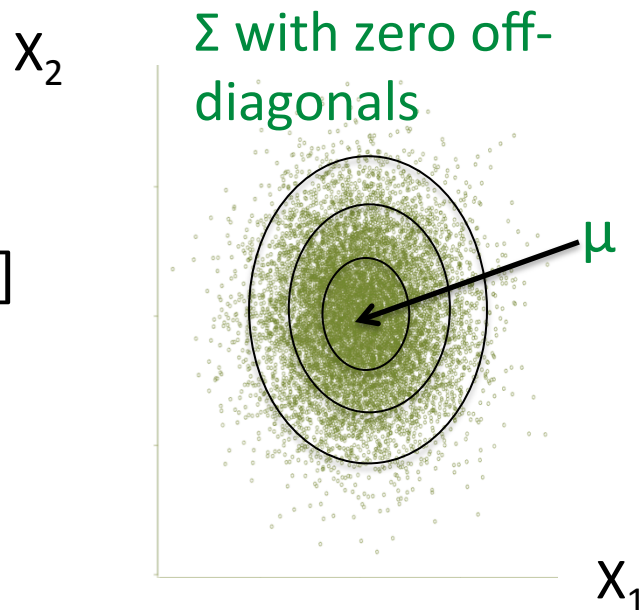
Linear instead of Quadratic in d!

Independent Gaussians

Equivalent to
assuming

$$\Sigma_y = \begin{bmatrix} \sigma_1^{2(y)} & 0 & 0 & 0 \\ 0 & \sigma_2^{2(y)} & 0 & 0 \\ & & \dots & \\ 0 & 0 & 0 & \sigma_d^{2(y)} \end{bmatrix}$$

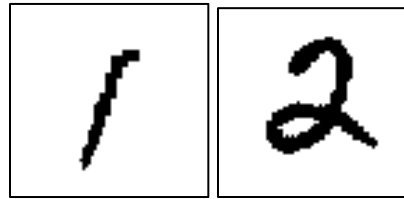
d=2
 $X = [X_1; X_2]$



Handwritten digit recognition (discrete features)

Training Data:

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



... n black-white (1/0)
images with
d pixels

Y

1

2

... n labels

How many parameters?

Class probability $P(Y = y) = p_y$ for all y **K-1 if K labels**

May not
hold

Class conditional distribution of features (using Naïve Bayes assumption)

$P(X_i = x_i | Y = y)$ – one probability value for each y, pixel i **Kd**

Linear instead of Exponential in d!

Naïve Bayes Classifier

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

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

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

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

How to learn parameters from data?

MLE, MAP

(Discrete case)

Learning parameters in distributions

$$P(Y = \bullet) = \theta$$

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

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

How do you learn that?

Data =



Answer: 3/5

Why??

Bernoulli distribution

Data, $D =$



- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$
- Flips are **i.i.d.**:
 - **Independent** events
 - **Identically distributed** according to Bernoulli distribution

Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation (MLE)

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

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

MLE of probability of head:

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

"Frequency of heads"

Multinomial distribution

Data, D = rolls of a dice



- $P(1) = p_1, P(2) = p_2, \dots, P(6) = p_6$ $p_1 + \dots + p_6 = 1$
- Rolls are **i.i.d.**:
 - **Independent** events
 - **Identically distributed** according to Multinomial(θ) distribution where $\theta = \{p_1, p_2, \dots, p_6\}$

Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D | \theta)$$

MLE of probability of rolls:

$$\hat{\theta}_{MLE} = \hat{p}_{1,MLE}, \dots, \hat{p}_{6,MLE}$$

$$\hat{p}_{y,MLE} = \frac{\alpha_y \leftarrow \text{Rolls that turn up } y}{\sum_y \alpha_y \leftarrow \text{Total number of rolls}}$$

"Frequency of roll y "

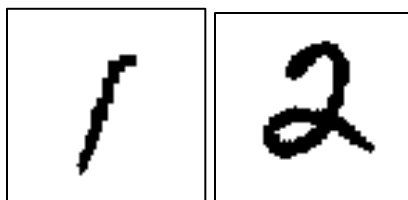
Back to Naïve Bayes

Naïve Bayes with discrete features

Training Data:

Each image represented as a vector of **d binary features** (black 1 or white 0)

Input, X



... n **black-white**
images

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

Label, Y

1

2

... n labels

Discrete Naïve Bayes model:

$P(Y = y) = p_y$ for all y in 0, 1, 2, ..., 9 p_0, p_1, \dots, p_9 (sum to 1)

$P(X_i = x_i | Y = y)$ - one probability value for each y , pixel i

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$
- Maximum Likelihood Estimates

- For Class probability

$$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

- For class conditional distribution

$$\frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \dots, x_d)$

$$Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

Issues with Naïve Bayes

- **Issue 1:** Usually, features are not conditionally independent:

$$P(X_1 \dots X_d | Y) \neq \prod_i P(X_i | Y)$$

Nonetheless, NB is the single most used classifier particularly when data is limited, works well

- **Issue 2:** Typically use MAP estimates instead of MLE since insufficient data may cause MLE to be zero.

Insufficient data for MLE

- What if you never see a training instance where $X_1=a$ when $Y=b$?
 - e.g., $b=\{\text{SpamEmail}\}$, $a=\{\text{'Earn'}\}$
 - $P(X_1=a \mid Y=b) = 0$
- Thus, no matter what the values X_2, \dots, X_d take:

$$\hat{P}(X_1 = a, X_2 \dots X_n | Y) = \hat{P}(X_1 = a | Y) \prod_{i=2}^d \hat{P}(X_i | Y) = 0$$

- What now???