# 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 $P(Y)$

  Class conditional distribution of features is Gaussian $P(X|Y)$

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

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

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

- Class distribution:

K classes

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

**K-1 if K labels** $\qquad$ K-1

- Class conditional distribution of features:

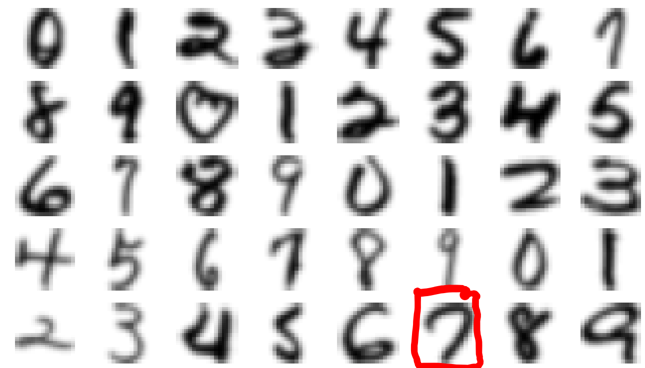
$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y $\qquad$ $\mu_y$ – d-dim vector

$\Sigma_y$ - dxd matrix

**$Kd + Kd(d+1)/2 = O(Kd^2)$ if d features** $\quad O(K(d+d^2))$ $\quad \Sigma_y(i,j) = \Sigma_y(j,i)$

**Quadratic in dimension d! If d = 256x256 pixels, ~ 13 billion parameters!** $\quad K\left(d + \frac{d(d+1)}{2}\right)$

# 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

$p_0, p_1, \dots, p_9$ (sum to 1)     K-1

**K-1 if K labels**

$X = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$ d-dim

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!**

# Naïve Bayes Classifier

$X_i^{(j)}$

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

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

$$P(\underbrace{X_1, X_2}_{X}|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

  - More generally:

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

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

- 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:

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

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

$= \arg\max_y P(y \mid x_1 \ldots x_d)$

- 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)?

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ $\quad E[X] = \begin{bmatrix} EX_1 \\ \vdots \\ EX_d \end{bmatrix}$

$\boxed{P(X|Y)} = \prod_{i=1}^{d} \underbrace{P(x_i|Y)}_{N(\mu_y, \sigma_y^2)}$

$N\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 \ddots & \\ 0 & & \sigma_K^2 \end{bmatrix}\right)$ $K \cdot 2d$

A. K-1, Kd

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

C. K-1, Kd

→ 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)?

$P(X|Y) = \prod_{i=1}^{d} P(X_i|Y)$ $\quad X_i \in \{0,1\}$

A. K-1, $K2^d$

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

C. K-1, Kd

D. K-1, 2Kd

$Kd \cdot 1$

vs.

$K2^d$

# 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]

Pairwise classification accuracy: 85%

$\mu_{people}$ People words


Animal words $\mu_{animal}$