CS 4824/ECE 4424: Gaussian Naïve Bayes

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Naïve Bayes in a Nutshell

- Bayes Rule:
  - $$P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n | Y = y_j)}$$
- Assuming conditional independence among X_i's

conditional independence

1 + 2n

  - $$P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)\prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i | Y = y_j)}$$
- How to pick the most probable Y for $X^{New} = \langle X_1, X_2, \ldots, X_n \rangle$?

$$Y^{New} \leftarrow \arg\max_{y_k} \boxed{P(Y = y_k)\prod_i P(X_i^{New} | Y = y_k)}$$

Naïve Bayes algorithm - discrete X_i

- Train Naïve Bayes (examples)
  - For each value $\boxed{y_k}$
    - Estimate $\pi_k = P(Y = y_k)$
    - For each value x_{ij} of each attribute $\boxed{X_i}$
      - Estimate $\underline{\theta_{ijk}} = P(X = \boxed{x_{ij}} \mid Y = y_k)$
    - *Note: Prob. must sum to 1 so we only need to estimate n-1 of these*

- Classify X^{New}
  - $Y^{New} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{New} \mid Y = y_k)$
  - $Y^{New} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$

inference or prediction

Another way to view Naïve Bayes (boolean Y)

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

- Decision rule:

$$1 \gtrless \frac{P(Y=1 \mid X_1 \cdots X_n)}{P(Y=0 \mid X_1 \cdots X_n)} = \frac{P(Y=1)\prod_i P(X_i \mid Y=1)}{P(Y=0)\prod_i P(X_i \mid Y=0)}$$

$$0 \gtrless \ln \frac{P(Y=1)}{P(Y=0)} \frac{\prod_i P(X_i \mid Y=1)}{\prod_i P(X_i \mid Y=0)}$$

$$\boxed{0 \gtrless \ln \frac{P(Y=1)}{P(Y=0)} + \sum_i \ln \frac{P(X_i \mid Y=1)}{P(X_i \mid Y=0)}}$$

linear sum of a prior term and conditional prob. terms

What if we have continuous X_i

- For example, image classification
  - X_i is the ith pixel, Y = mental state

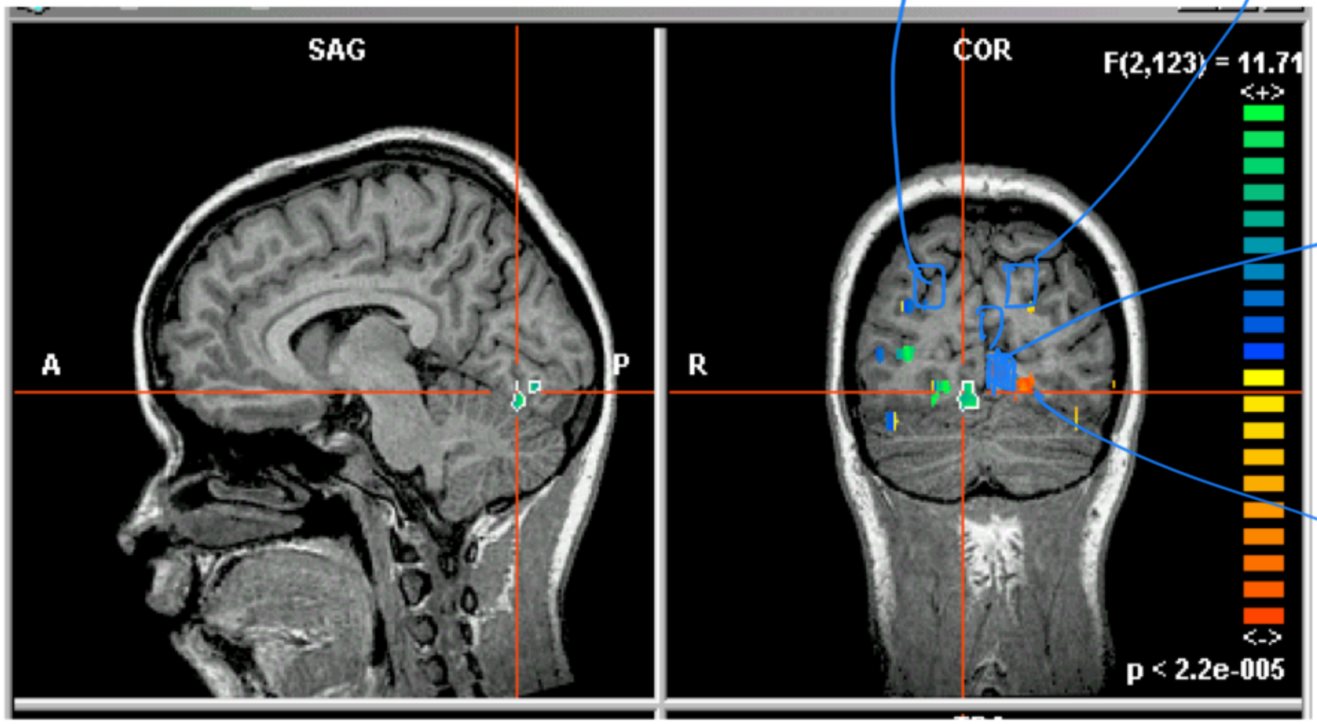
x_i

x_i

x_5

Normal distribution

x_7



  - We still have

$$P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)\prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i | Y = y_j)}$$

$P(X)$

all possible values of y

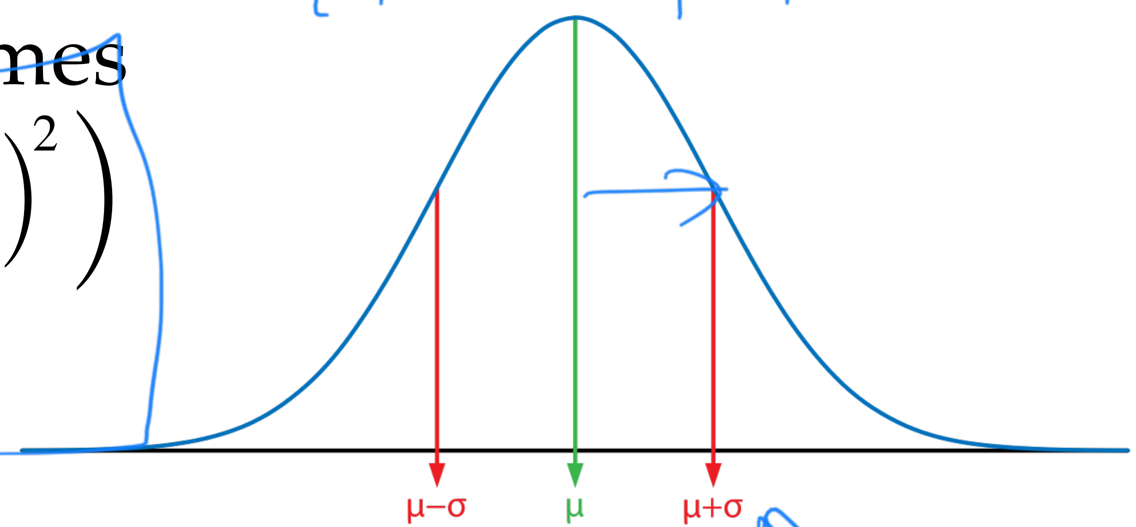
  - **How to represent $P(X_i | Y)$?**

What if we have continuous X_i

- For example, image classification
  - Xi is the ith pixel, Y = mental state
  - Gaussian Naïve Bayes (GNB) assumes

$$\boxed{P(X_i \mid Y)} = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2\right)}$$

k is the class label
i is the feature



means

var.

- Sometimes assume σ_{ik}
  - is independent of Y (i.e., σ_i)
  - is independent of X_i (i.e., σ_k)
  - or both (i.e., σ)

What are the implications of these assumptions?

Gaussian Naïve Bayes algorithm: continuous X_i but discrete Y

- Train Naïve Bayes (examples)
  - For each value y_k
    - Estimate $\pi_k = P(Y = y_k)$
    - For each value x_{ij} of each attribute X_i
      - Estimate class conditional μ_{ik} and variance σ_{ik}

    - *Note: Prob. must sum to 1 so we only need to estimate n-1 of these*

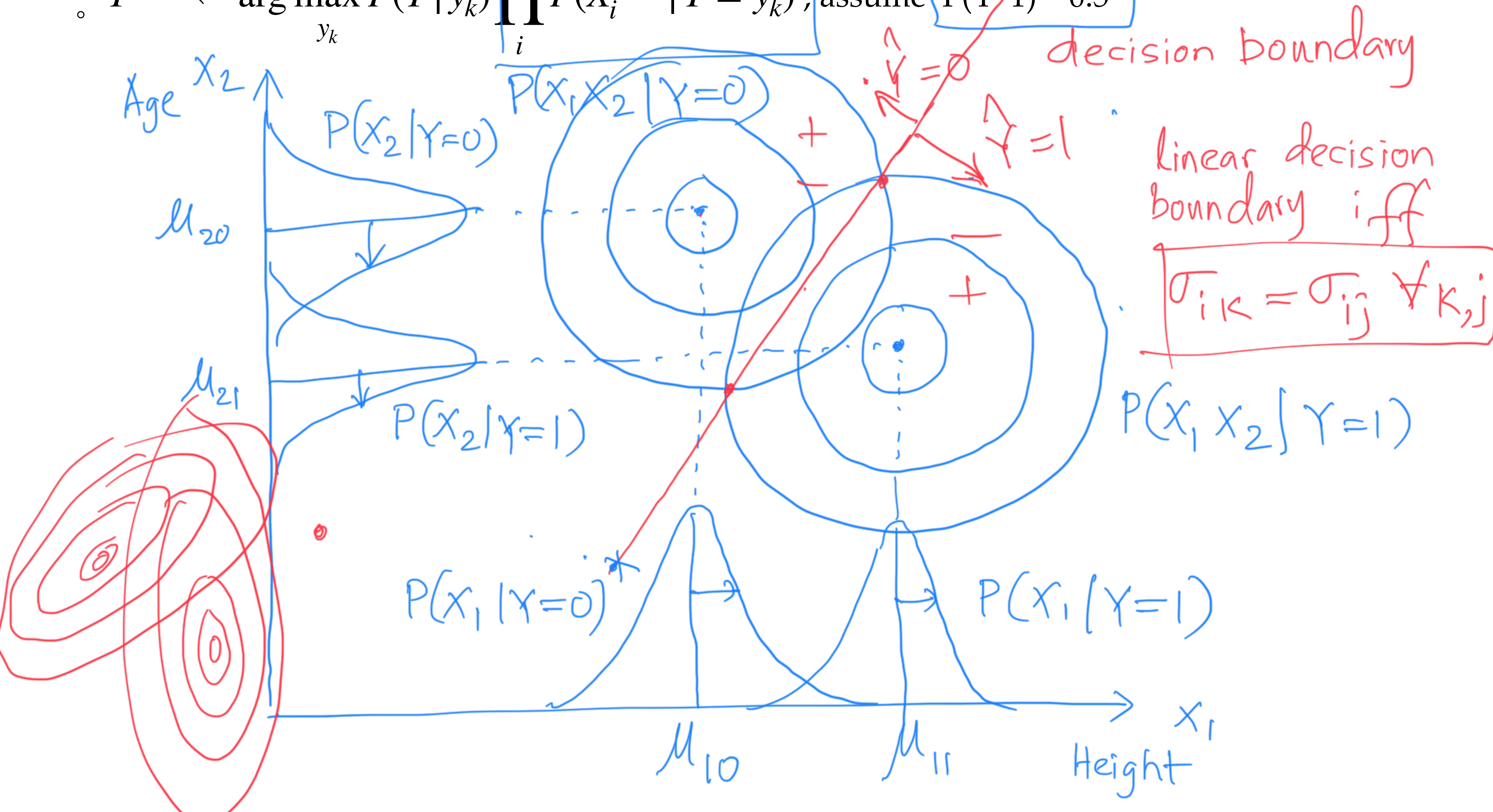
- Classify X^{New}
  - $$Y^{New} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{New} \mid Y = y_k)$$
  - $$Y^{New} \leftarrow \arg\max_{y_k} \pi_k \prod_i \mathcal{N}(X_i^{New}, \mu_{ik}, \sigma_{ik})$$

Estimating parameters: continuous X_i but discrete Y

- MLE
  - $$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$
  - $$\sigma_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Gaussian Naïve Bayes - decision surface

- Assume Y=PlayBasketball (boolean) X1=Height X2=Age
  - $Y^{New} \leftarrow \arg\max_{y_k} P(Y \mid y_k) \prod_i P(X_i^{New} \mid Y = y_k)$; assume P(Y=1) = 0.5



What is the minimum possible error?

- Best case:
  - Conditional independence assumption is satisfied
  - We can perfectly estimate P(Y), P(X|Y) (i.e. infinite training data)

But…

- Naïve Bayes allows estimating P(Y|X) by learning P(Y) and P(X|Y)

- Why not learn P(Y|X) directly?