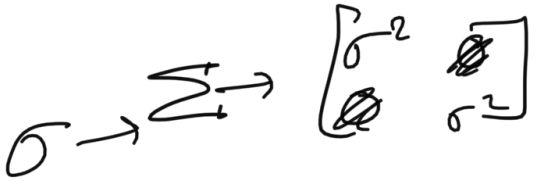
CSE 575: Statistical Machine Learning (Spring 2021)
Instructor: Nupur Thakur

Supervised Learning

Table of contents

1. MLE solution for variance
2. Naive Bayes
3. Logistic Regression

MLE - Example 1 (Continued)



Given *n* i.i.d. samples $\{x_i\}$ from the 1-D normal distribution $N(\mu, \sigma^2)$, find the MLE for μ and σ^2.

$$\log P(D|\mu,\sigma) = \underbrace{-n\log(\sigma\sqrt{2\pi})}_{-\frac{n}{2}\log(2\pi\sigma^2)} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\sigma}\log P(D|\mu,\sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}$$

Equate it 0,

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

MLE - Example 1 (Continued)

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} = \frac{n}{1}$$

$$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$ — Biased

Bias $= E(\hat{\sigma}_{MLE}) - \sigma^2 = 0$ — Unbiased

Naive Bayes

$P(Y|X)$ $y=3$

$P(X_1, X_2, X_3|Y)$ $Y=0$ $Y=1$ 2^3 2^3

- The "naive" conditional independence assumption: each feature is (conditionally) independent of every other feature, given the label, i.e., $p(x_i | \{x_j \text{ for any } j \neq i\}, y) = p(x_i|y)$
- How does this simplify?
  - Consider the previous example again: d-dimensional binary features, and y is also binary.

No. of probabilities $P(X_1, X_2, X_3|Y) = 2^{3+1}$

Bayesian $\Rightarrow$ d-dimensional $X \rightarrow 2 \cdot 2^{d+1} = 4d$

$P(X_1, X_2, X_3|Y) = P(X_1|Y) \cdot P(X_2|Y) \cdot P(X_3|Y)$

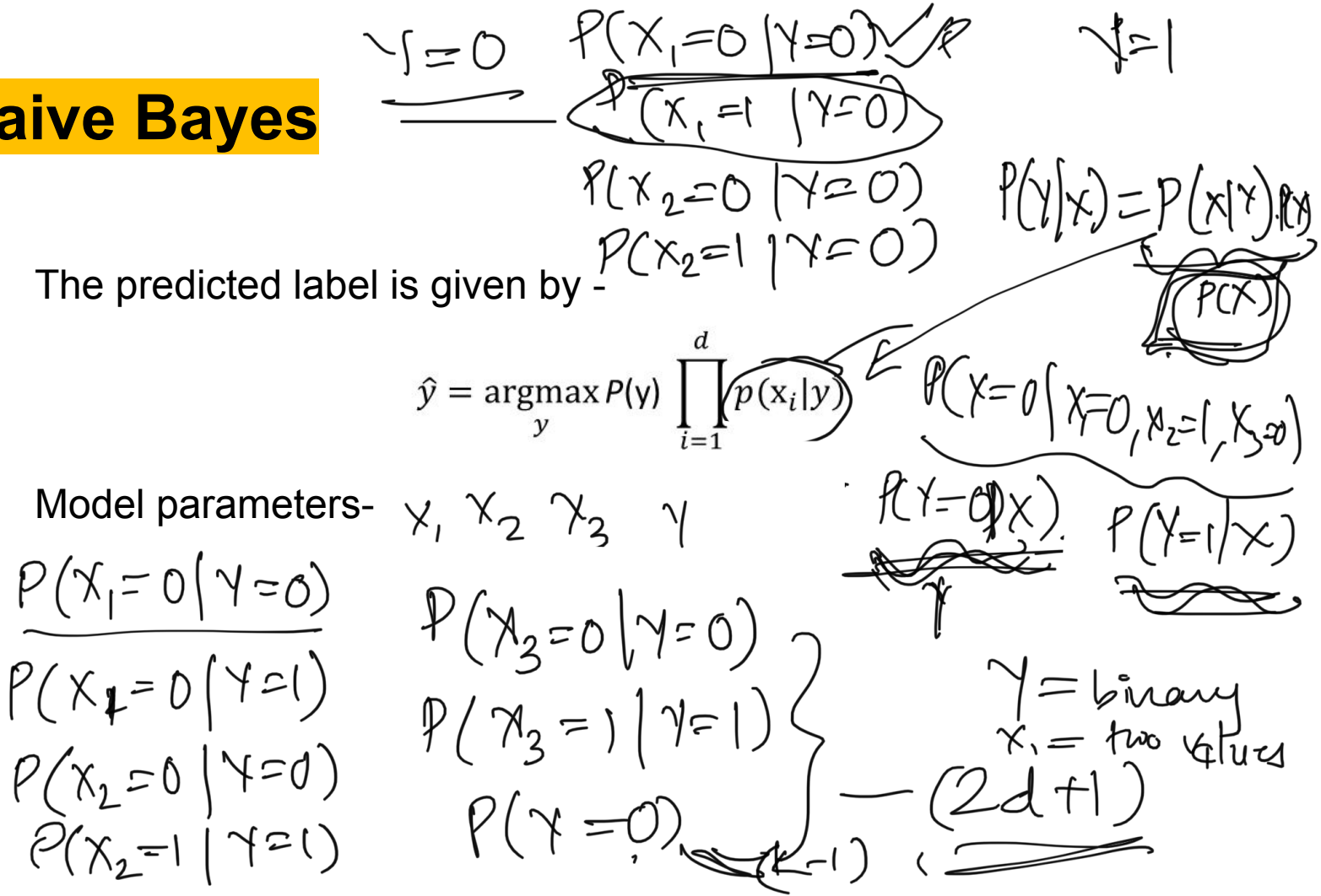
$Y=0 \rightarrow P(X_1=0|Y=0) \cdot P(X_1=1|Y=0)$ $6+6=12$

Naive Bayes

- The predicted label is given by -

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y) \prod_{i=1}^{d} p(x_i|y)$$

- Model parameters-



Naive Bayes - Example

X						Y
x_1	x_2	x_3	x_4	x_5	x_6	
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$P(Y = Yes) = \frac{3}{4}$

$P(X_1 = Sunny \mid Y = Yes) = \frac{3}{3}$

$(2d+1) = 13$ independent parameter

$P(X|Y)$

Logistic Regression

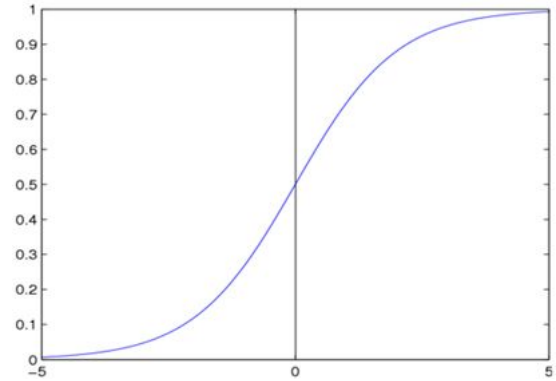
$w^T x \geq 0$

- Training set: n labelled samples <x(i), y(i)>
- Use the logistic function for modeling P(y|x), considering only the case of $y \in \{0,1\}$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)} = 1 - \sigma(w^T x)$$

$w^T x$

$$P(y=1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)} = \sigma(w^T x)$$



Sigmoid $\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}$

Logistic Regression

$w \rightarrow$ model parameter

- Model parameters - $P(Y|X) = \left[1 - \sigma(w^T x^{(i)})\right]^{1-y} \left[\sigma(w^T x)\right]^{y}$

$$P(y^{(1)}, y^{(2)}, \ldots y^{(n)} | x^{(1)} x^{(2)} \ldots x^{(n)}) = \prod_{i=1}^{n} \left[1 - \sigma(w^T x^{(i)})\right]^{(1-y^{(i)})} \left[\sigma(w^T x^{(i)})\right]^{y^{(i)}}$$

$$\text{argmax}\ l(w) = \sum_{i=1}^{n} \left[1 - y^{(i)}\right] \log\left[1 - \sigma(w^T (x^{(i)}))\right] + y^{(i)} \log\left[\sigma(w^T x^{(i)})\right]$$

Gradient Ascent

Update Equation —

$$w_{t+1} = w_t + \eta \nabla_w L(w)$$

- w_{t+1}: Updated weights
- w_t: Current weight
- η: Learning rate
- $\nabla_w L(w)$: Gradient of the Loss function

Descent —

Gradient Ascent

Questions?