Naïve Bayes

INSTRUCTOR: HONGJIE CHEN

MAY 26TH 2022

Review $P(Y|X)$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

Number of Parameters

- Consider a joint probability distrbution with 50 boolean features
  - $X = [X_1, X_2, \ldots, X_{50}]$
  - $P(Y \mid X_1, X_2, \ldots, X_{50})$, suppose Y is a boolean R.V.
  - How many parameters do we need to estimate?

- Can we leverage Bayes' theorem?

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Let's set $n = 50$

Number of Parameters with Bayes' theorem

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
- How many parameters to estimate $P(X_1, X_2, \ldots, X_n | Y)$?

- How many parameters to estimate $P(Y)$

Number of Parameters with Bayes' theorem

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
- How many parameters to estimate $P(X_1, X_2, \ldots, X_n | Y)$?

  $2(2^n - 1)$

- How many parameters to estimate $P(Y)$

  1

Review: Independence and Conditional Independence

- Independence: $P(X_1, X_2) = P(X_1)P(X_2)$
- Conditional Independence:
  $P(X_1, X_2 \mid Y) = P(X_1 \mid Y)P(X_2 \mid Y)$
- Extend to multiple variables:
  $P(X_1, X_2, \cdots, X_n \mid Y) = P(X_1 \mid Y)P(X_2 \mid Y)\cdots P(X_n \mid Y)$

Naïve Bayes Backbone

- $P(X_1, X_2, \cdots, X_n \mid Y) = \prod_i P(X_i \mid Y)$
  - $\forall i \neq j$, X_i and X_j are conditionally independent given Y.

- If A and B are conditionally independent given C, then
  - $P(A \mid B, C) = P(A \mid C)$

Proof

Incorporate Conditional Independence

- Recall $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$, assume Boolean.
- Number of parameters to estimate $P(X_1, X_2, \ldots, X_n | Y)$?
  - $2 * (2^n - 1)$
- Number of parameters to estimate $P(Y)$
  - 1
- Incoroprate conditional independence

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

Law of total probability

Parameters with Conditional Independence

- $$P(Y = y_k \mid X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n \mid Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n \mid Y = y_j)} = \frac{P(Y = y_k)\prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i \mid Y = y_j)}$$
- Assume Boolean again
- $P(Y)$: 1
- $P(X_i \mid Y)$: $2 * n$

Training and Testing Naïve Bayes

- $$P(Y = y_k \mid X_1, \ldots, X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}$$
  - Estimate $\pi_k = P(Y = y_k)$
  - For each value x_{ij} of each attribute X_i
    - Estimate $\theta_{ijk} = P(X_i = x_{ij} \mid Y = y_k)$
- When classifying X^{New}

$$Y^{New} \leftarrow \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{New} \mid Y = y_k)$$

$$= \text{argmax}_{y_k} \pi_k \prod_i \theta_{ijk}$$

A Real Life Example: 20 newsgroup dataset

- Link
- Having a set of 18000 articles
- Each article contains several words, X
- Each article belongs to one of the 20 categories, Y

  - Estimate $\pi_k = P(Y = y_k), k = 1,2,\ldots,20$
  - Estimate $\theta_{ijk} = P(X_i = x_{ij} \mid Y = y_k)$

    - X_i is the i(th) article, x_{ij} is the j(th) word of X_i

Log Trick

$$Y^{New} \leftarrow \text{argmax}_{y_k} \pi_k \prod_i \theta_{ijk}$$

$$Y^{New} = \text{argmax}_{y_k} \pi_k \ln \prod_i \theta_{ijk}$$

$$= \text{argmax}_{y_k} \pi_k \sum_i \ln \theta_{ijk}$$

Essentially likelihood and log likelihood lead to the same result

Parameter Estimation with MLE

- Maximum Likelihood Estimation

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \,|\, Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Concern - I

- A strong assumption: $P(X_1, X_2, \cdots, X_n \mid Y) = \prod_i P(X_i \mid Y)$
  - X_i are conditionally independent given Y
- Imagine we have two features $X_i = X_k$, what's the effect?

Concern - II

- MLE for $P(X_i \mid Y)$ may be zero

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

We are calculating $\prod_{i=1}^{n} P(X_i \mid Y)$, what if for c: $P(X_c \mid Y) = 0$

Parameter Estimation with MAP

- Maximum A Posteriori
- What should be the prior?

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \,|\, Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Beta and Dirichlet Prior

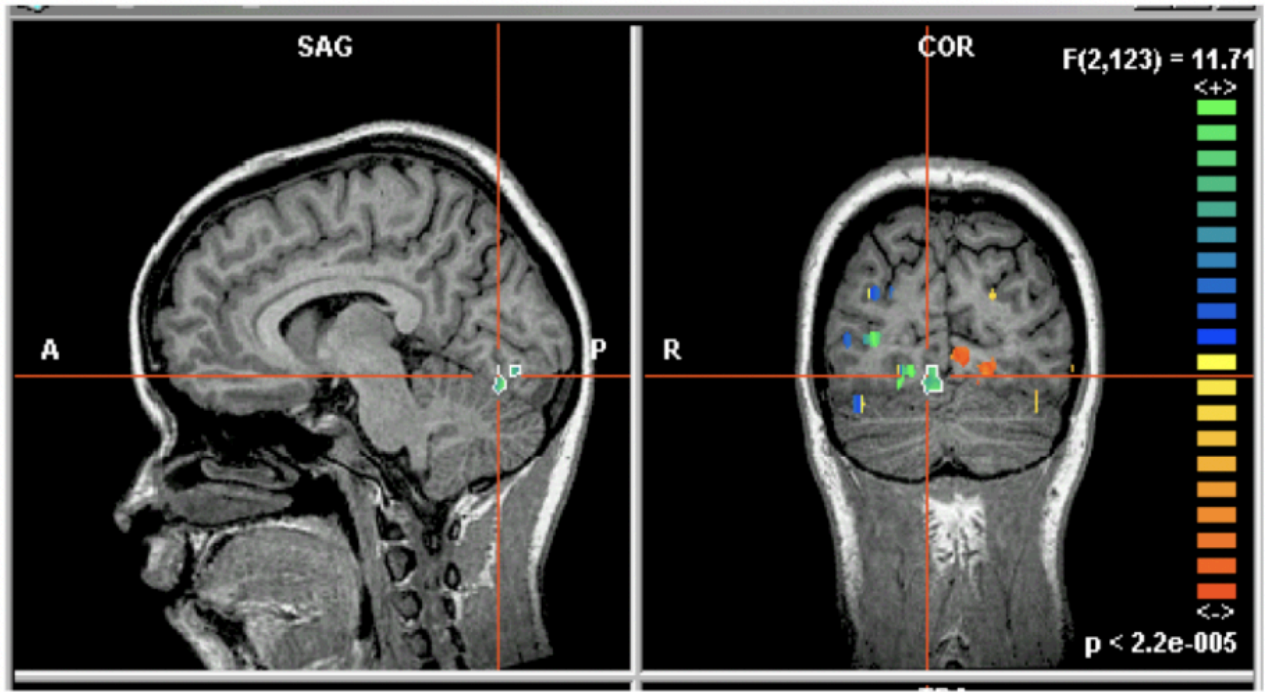
- Maximum A Posteriori

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \,|\, Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

Continuous X

- Let's move from discrete X to continuous X.



- How to represent $P(X_i \mid Y)$

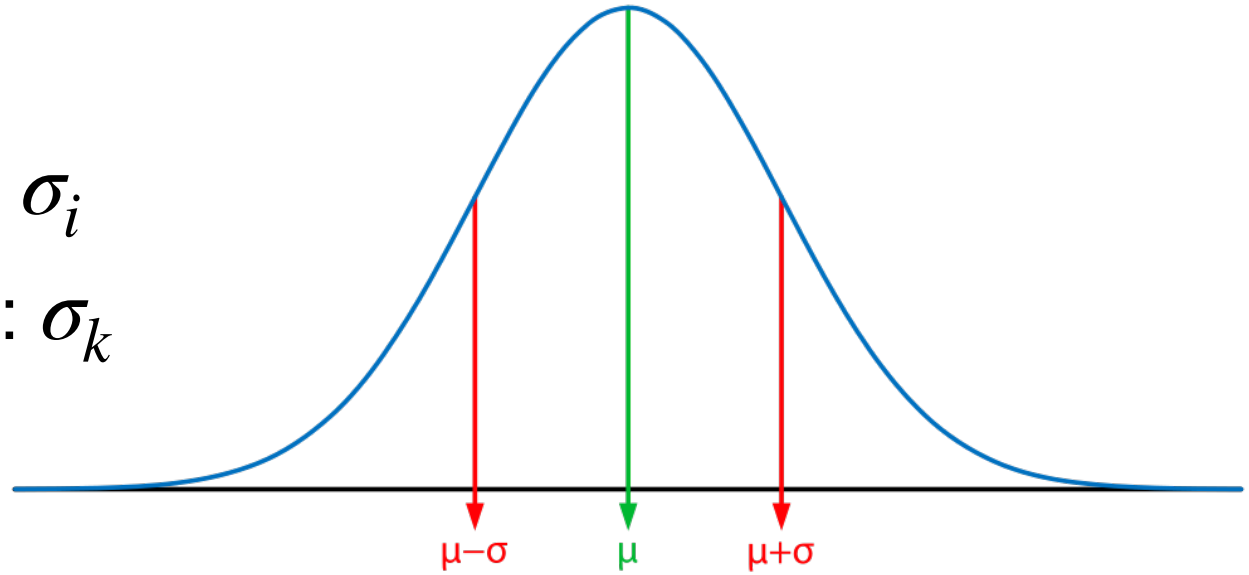
Figure credit: link

Gaussian Naïve Bayes (GNB)

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

Sometimes assume σ_{ik}

- is indendendent of Y: σ_i
- is indendendent of X_i: σ_k
- Or both: σ



Gaussian Naïve Bayes Training and Testing

- $P(Y = y_k \mid X_1, \ldots, X_n) = = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}$ **Discrete Y**
  - 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}
- Classify X^{New}

$$Y^{New} \leftarrow \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{New} \mid Y = y_k)$$

$$= \text{argmax}_{y_k} \pi_k \prod_i \mathcal{N}(X_i^{New}, \mu_{ik}, \sigma_{ik})$$

Gaussian Naïve Bayes with MLE

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

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

$\delta(x) = 1$, when x is True