

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2026



HAVERFORD
COLLEGE

Admin

- **Lab 5** posted (due Tuesday after break)
- **Lab 3** grades & feedback posted on Moodle
- **Mid-semester feedback form** (link on Piazza)

Outline for today

- Intro to probability
 - Bayes' Rule
- Intro to Bayesian models
- Naïve Bayes algorithm

Outline for today

- Intro to probability
 - Bayes' Rule
- Intro to Bayesian models
- Naïve Bayes algorithm

Intro to Probability

- The **probability** of an **event** e has a number of epistemological interpretations
- Assuming we have **data**, we can count the number of times e occurs in the dataset to estimate the probability of e , $P(e)$

$$P(e) = \frac{\text{count}(e)}{\text{count}(\text{all events})}$$

- If we put all events a bag, shake it up, and choose one at random (called **sampling**), how likely are we to get e ?

Intro to Probability

Probability Axioms

1. Probabilities of events must be no less than 0. $P(e) \geq 0$ for all e .

2. The sum of all probabilities in a distribution must sum to 1.

That is, $P(e_1) + P(e_2) + \cdots + P(e_n) = 1$

Or, more succinctly, $\sum_{e \in E} P(e) = 1$

Intro to Probability

Joint Probability

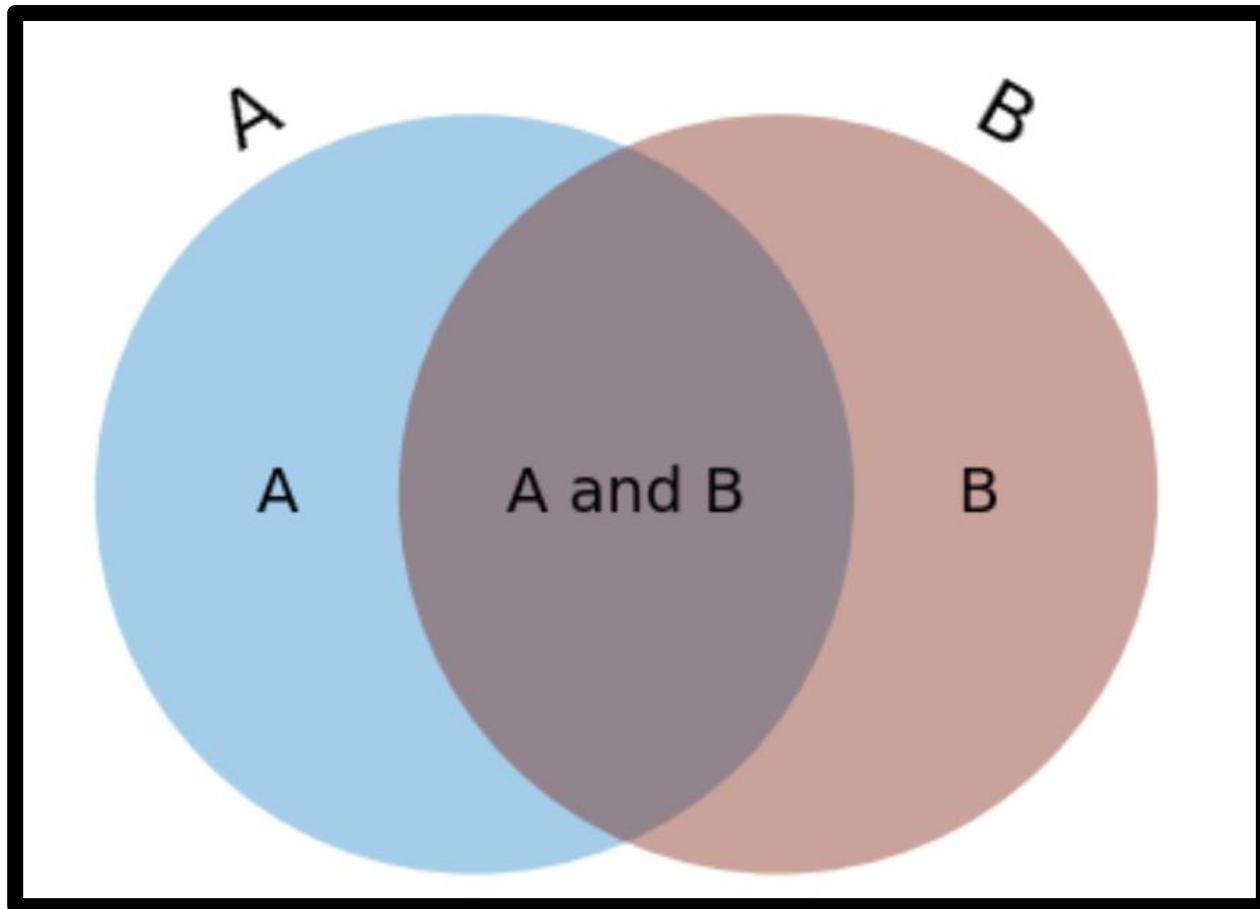
The probability that two independent events e_1 and e_2 both occur is given by their product.

$$P(e_1 \wedge e_2) = P(e_1 \cap e_2) = P(e_1)P(e_2) \text{ when } e_1 \cap e_2 = \emptyset$$

- Intuitively, think of every probability as a *scaling* factor.
- You can think of a probability as the fraction of the probability space occupied by an event e_1
 - $P(e_1 \wedge e_2)$ is the fraction of e_1 's probability space wherein e_2 also occurs.
 - So, if $P(e_1) = \frac{1}{2}$ and $P(e_2) = \frac{1}{3}$, then $P(e_2, e_1)$ is a third of a half of the probability space or $\frac{1}{3} \times \frac{1}{2}$

Intro to Probability

Joint Probability



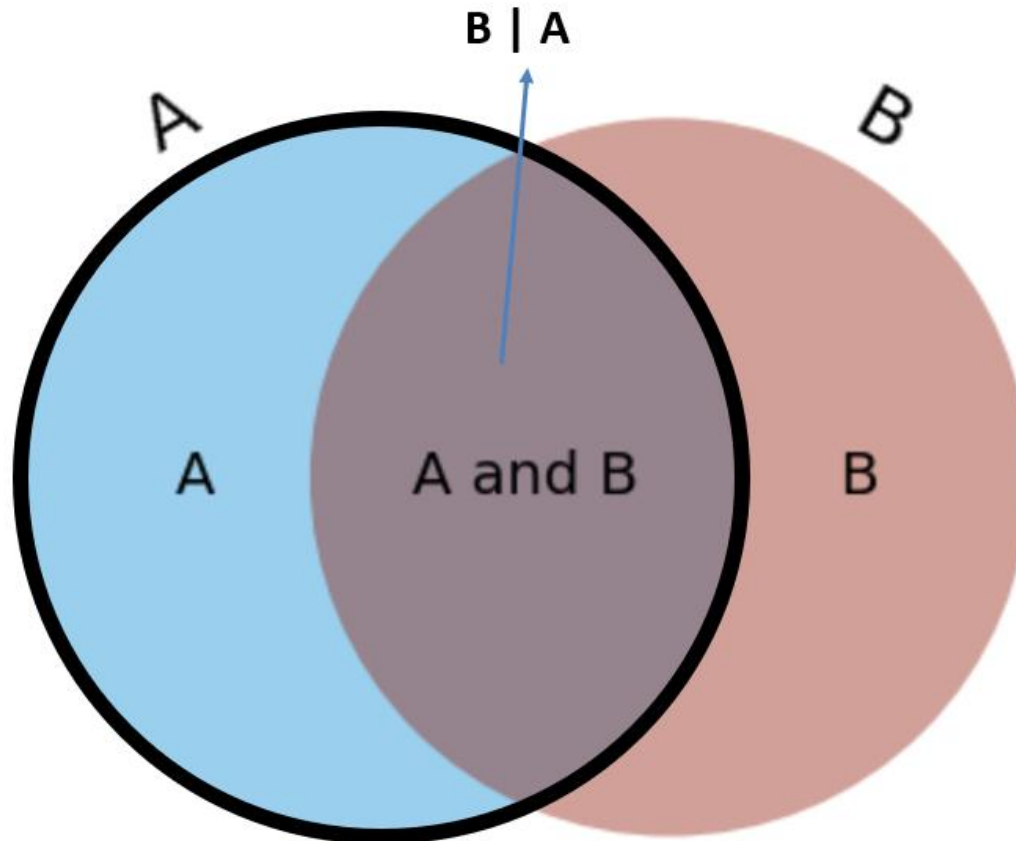
Intro to Probability

Conditional Probability

- A conditional probability is the probability that one event occurs given that we take another for granted.
- The probability of e_2 given e_1 is $P(e_2 \mid e_1)$
- This is the probability that e_2 will occur given that we take for granted that e_1 occurs.

Intro to Probability

Conditional Probability



Intro to Probability

Marginal Probability Distributions

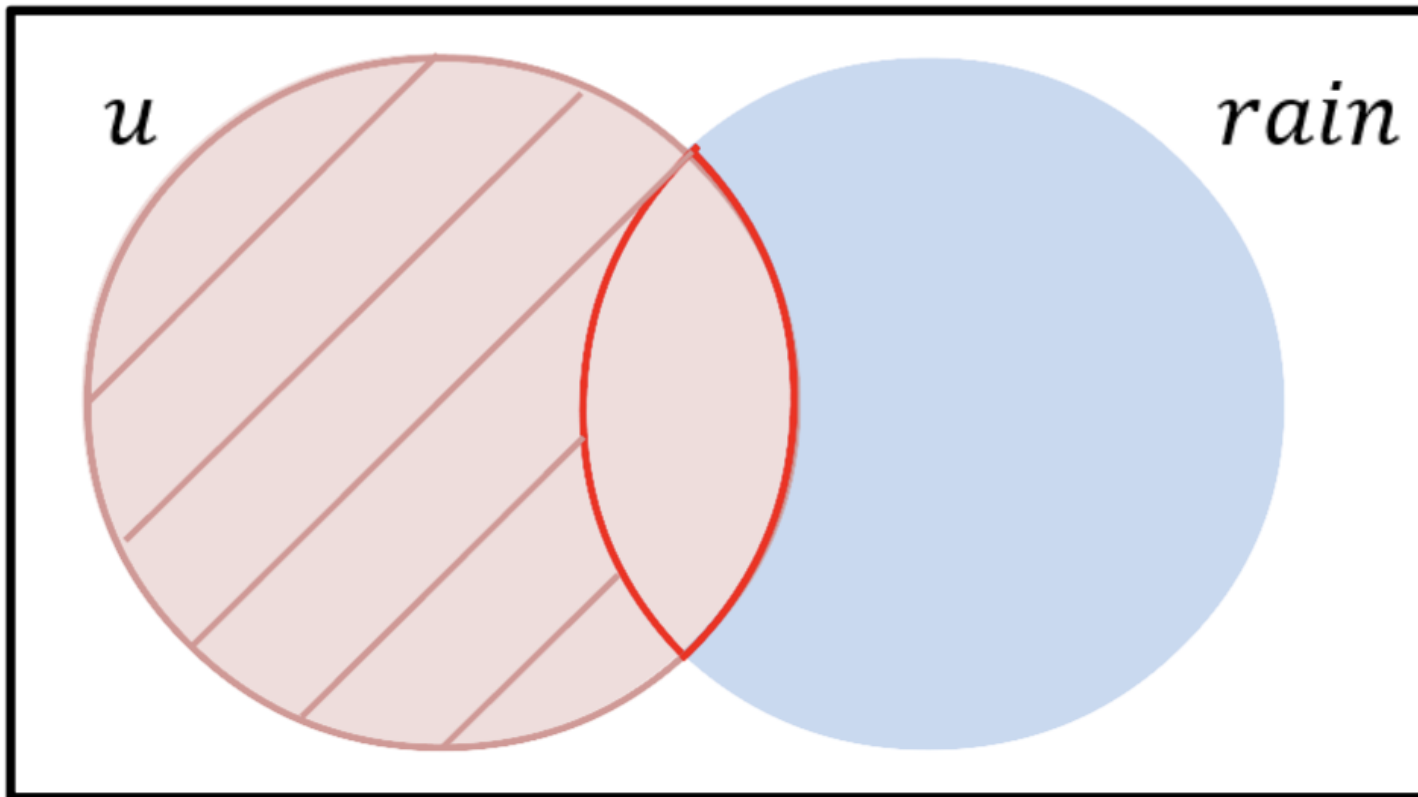
Given a discrete joint probability distribution function $P(X, Y)$, how would we find $P(X)$?

- “Marginalize out” the Y (sum over all $y \in Y$)
- Discrete Case: $p(x) = \sum_{y \in Y} P(x, y)$
- Continuous Case: $p(x) = \int p(x, y) dy$

Intro to Probability

Marginal Probability Distributions

Example: $P(u) = P(u, rain) + P(u, \overline{rain})$



Example

- $R = \text{rain}$, $U = \text{umbrella}$
- If $P(R) = 20\%$ and $P(R,U) = 15\%$. What is $P(U|R)$?

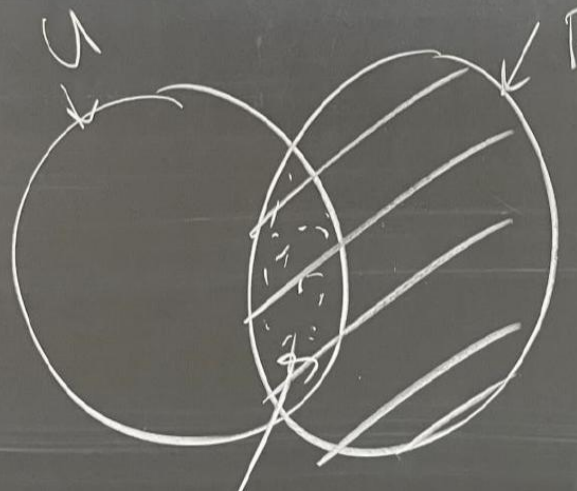
Bayes' Theorem

- $P(A, B) = P(A|B)P(B)$
- $P(A, B) = P(B|A)P(A)$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example

If $P(R) = 20\%$ and $P(R, U) = 15\%$,
what is $P(U|R)$?



$\boxed{\text{Intersection}} = 0.15$

$\boxed{\text{Region of R outside U}} = 0.2$

$$P(U|R) = \frac{P(R, U)}{P(R)}$$
$$= \frac{15}{20} = \boxed{.75}$$

Sum over all events
should be 1

Example
Y or N email

$$P(\text{spam} | \text{words}) =$$

+ want to compute
"posterior"

$$\frac{P(\text{spam}, \text{words})}{P(\text{words})}$$

"X" "data"
very difficult!

$$P(\text{spam}, \text{words})$$

$$= \frac{P(\text{words}, \text{spam}) + P(\text{words}, \overline{\text{spam}})}{P(\text{spam}) + P(\overline{\text{spam}})}$$

"prior"

$$= \frac{P(\text{spam}) P(\text{words} | \text{spam})}{P(\text{spam}) P(\text{words} | \text{spam}) + P(\overline{\text{spam}}) P(\text{words} | \overline{\text{spam}})}$$

"likelihood"
(generative)

$$P(\text{spam}) P(\text{words} | \text{spam}) + P(\overline{\text{spam}}) P(\text{words} | \overline{\text{spam}})$$

"evidence"

"prior"

Handout 10

Handout 10

$$\begin{aligned} P(D|pos) &= \frac{P(D)P(pos|D)}{P(pos)} \\ &= \frac{P(D)P(pos|D)}{P(pos, D) + P(pos, \bar{D})} \\ &= \frac{P(D)P(pos|D)}{P(D)P(pos|D) + P(\bar{D})P(pos|\bar{D})} \end{aligned}$$

Handout 10

$$p(\text{neg} | H) = 0.9$$

$$p(\text{neg} | H) + p(\text{pos} | H) = 1$$

↑ ~
or D

$$\frac{1}{100} \cdot \frac{9}{10}$$

$$\frac{1}{100} \cdot \frac{9}{10} + \frac{99}{100} \cdot \frac{1}{10}$$

$$\frac{9}{100}$$

$$= \frac{1}{12}$$

$$\approx \boxed{8\%}$$

Outline for today

- Intro to probability
 - Bayes' Rule
- Intro to Bayesian models
- Naïve Bayes algorithm

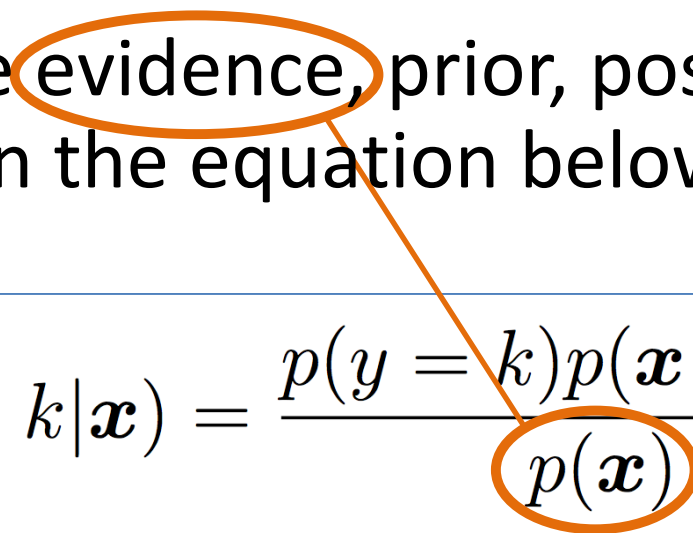
Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$


- Evidence:** this is the data (features) we actually observe, which we think will help us predict the outcome we're interested in

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Prior**: without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and **likelihood** in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Likelihood**: given an outcome, what is the probability of observing this set of features?

Components of a Bayesian Model

- Identify the evidence, prior, **posterior**, and likelihood in the equation below

$$p(y = k | \mathbf{x}) = \frac{p(y = k)p(\mathbf{x} | y = k)}{p(\mathbf{x})}$$

- Posterior**: this is the quantity we are actually interested in. **Given** the evidence, what is the probability of the outcome?

Examples

- Computing the probability an email message is **spam**, given the **words** of the email
- Another example: what is the probability of **Trisomy 21** (Down Syndrome), given the **amount of sequencing of each chromosome?**

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,...,n):

$$q_1, q_2, \dots, q_n = \vec{q}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,...,n):

$$q_1, q_2, \dots, q_n = \vec{q}$$

Goal:

$$\begin{aligned}\mathbb{P}(T_{21}|\vec{q}) &= \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q})} \\ &= \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q} | T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}\end{aligned}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,...,n):

$$q_1, q_2, \dots, q_n = \vec{q}$$

Goal:

$$\mathbb{P}(T_{21} | \vec{q}) = \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q})}$$

Prior probability of T_{21}

$$= \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q} | T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}$$

Prior:

$P(21)$

Maternal Age	Trisomy 21	All Trisomies
20	1 in 1,667	1 in 526
21	1 in 1,429	1 in 526
22	1 in 1,429	1 in 500
23	1 in 1,429	1 in 500
24	1 in 1,250	1 in 476
25	1 in 1,250	1 in 476
26	1 in 1,176	1 in 476
27	1 in 1,111	1 in 455
28	1 in 1,053	1 in 435
29	1 in 1,000	1 in 417
30	1 in 952	1 in 384
31	1 in 909	1 in 384
32	1 in 769	1 in 323
33	1 in 625	1 in 286
34	1 in 500	1 in 238
35	1 in 385	1 in 192
36	1 in 294	1 in 156
37	1 in 227	1 in 127
38	1 in 175	1 in 102
39	1 in 137	1 in 83
40	1 in 106	1 in 66
41	1 in 82	1 in 53
42	1 in 64	1 in 42
43	1 in 50	1 in 33
44	1 in 38	1 in 26
45	1 in 30	1 in 21
46	1 in 23	1 in 16
47	1 in 18	1 in 13
48	1 in 14	1 in 10
49	1 in 11	1 in 8

Independence

$$\begin{array}{l} P(A, B) = P(A)P(B) \\ \Downarrow \\ P(A|B)P(B) = P(A)P(B) \end{array} \left. \vphantom{\begin{array}{l} P(A, B) = P(A)P(B) \\ \Downarrow \\ P(A|B)P(B) = P(A)P(B) \end{array}} \right\} \text{not always true!}$$

Conditional Independence

$$P(A|B, C) = P(A|C)$$

“A is independent of B given C”

thunder rain lightning



Outline for today

- Intro to probability
 - Bayes' Rule
- Intro to Bayesian models
- Naïve Bayes algorithm

Real-world example of Naïve Bayes

[“A Comparison of Event Models for Naive Bayes Text Classification”](#) (6000+ citations!)

Goal: text classification (classify documents into topics based on the words as features)

95 topics (i.e., $K=95$)

Naïve Bayes

- Single example: $\vec{x} = [x_1, x_2, \dots, x_p]^T$
- Multi-class label: $y \in \{1, 2, \dots, K\}$
- Goal: Classification $\hat{y} = \operatorname{argmax}_{k=1, \dots, K} p(y = k | \vec{x})$

Bayesian Model

$$p(y = k | \vec{x}) = \frac{p(y = k)p(\vec{x} | y = k)}{p(\vec{x})}$$

can ignore

Naïve Bayes

$$p(\vec{x}|y = k) = p(\underbrace{x_1}_A, \underbrace{x_2, x_3, \dots, x_p}_B | y = k) \quad P(A,B)=P(B)P(A|B)$$

$$= p(\underbrace{x_2, x_3, \dots, x_p}_{\substack{C \\ D}} | y = k) p(\underbrace{x_1}_A | \underbrace{x_2, \dots, x_p}_B, y = k)$$

$$= p(\underbrace{x_3, \dots, x_p}_D | y = k) p(\underbrace{x_2}_C | \underbrace{x_3, \dots, x_p}_D, y = k) \\ p(x_1 | x_2, \dots, x_p, y = k)$$

Naïve Bayes assumption

Conditional Independence: “feature j is independent from all other features given label k ”

$$p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | x_1, y) \leftarrow \text{Bayes rule}$$

$x_1 = 4$ legs

$x_2 = \text{fur}$

$y = \text{cat}$

assume $p(x_2 | x_1, y) = p(x_2 | y)$

$$\Rightarrow p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | y)$$

Naïve Bayes

$$\begin{aligned} p(\vec{x}|y = k) &= p(x_p|y = k)p(x_{p-1}|y = k) \dots p(x_2|y = k) p(x_1|y = k) \\ &= \prod_{j=1}^p p(x_j|y = k) \end{aligned}$$

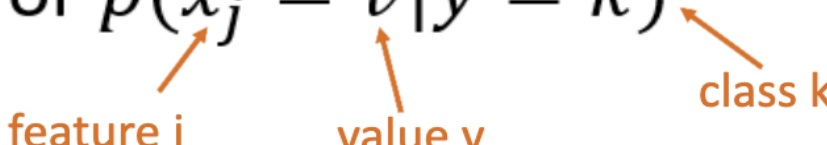
Naïve Bayes Model

$$p(y = k|\vec{x}) \propto p(y = k) \prod_{j=1}^p p(x_j|y = k)$$

proportional to

Obtaining $p(y = k)$ & $p(x_j | y = k)$

Estimate based on training data

- θ_k = estimate for $p(y = k)$
- $\theta_{k,j,v}$ = estimate for $p(x_j = v | y = k)$ 

Let N_k = # of examples with label k , we could define $\theta_k = \frac{N_k}{n}$

What happens if $N_k = 0$?

Laplace smoothing

- Technique to handle zero probability

$$\theta_k = \frac{N_k + 1}{n + K} ; \quad \sum \theta_k = \sum \frac{N_k + 1}{n + K} = \frac{1}{n + K} (n + K) = 1$$

 # of classes for y

- Similarly, let $N_{k,j,v}$ = number of examples with feature j = value v and class label k

$$\theta_{k,j,v} = \frac{N_{k,j,v} + 1}{N_k + |f_j|}$$

 # of feature values
for feature j