# 18s1: COMP9417 Machine Learning and Data Mining

**Lectures**: Supervised Learning – Classification
**Topic**: Questions from lectures
**Version**: with answers

**Last revision**: Sun Apr 8 17:25:20 AEST 2018

## Introduction

Some questions and exercises from the course lectures covering aspects of learning linear models (models "linear in the parameters") for classification tasks.

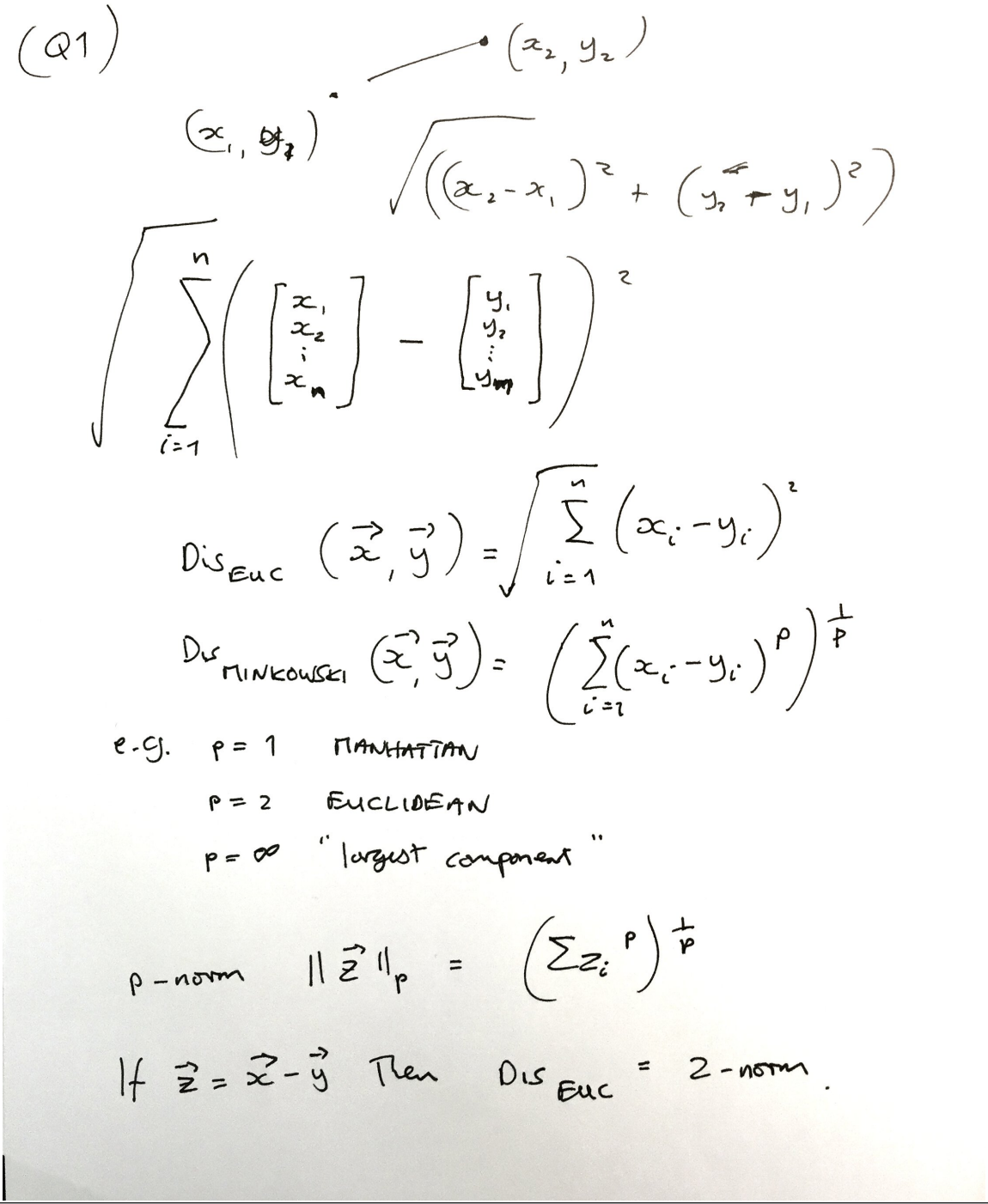
## Nearest-neighbour classifier

**Question 1** Which $p$-norm corresponds to *Euclidean* distance ? Express Euclidean distance in terms of this norm.

---

**Answer**

(Q1)

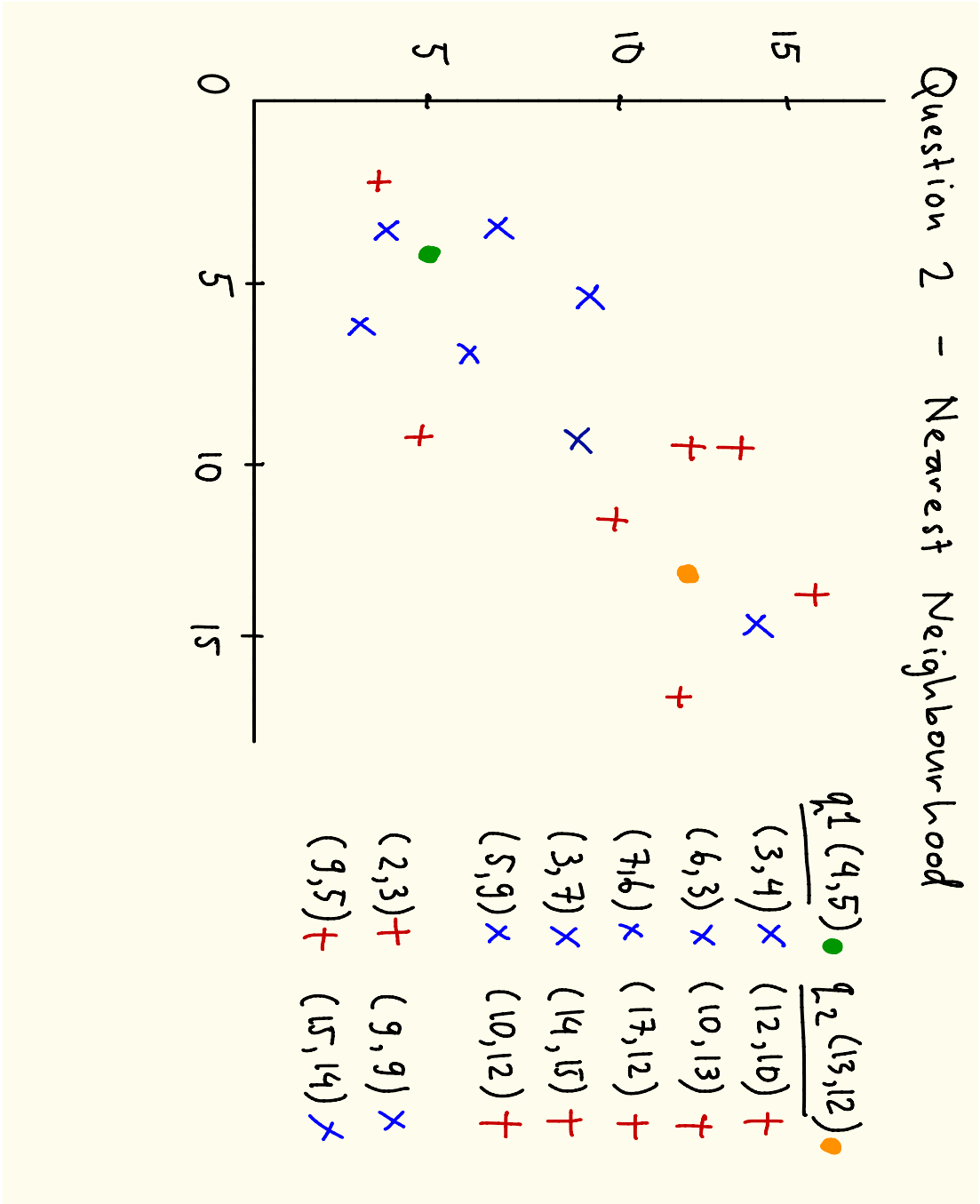



**Question 2** Construct a small two dimensional data set and show, for two different query points, how classification by $k$ nearest neighbour is affected by (a) the number of neighbours (i.e., the value

of $k$), and (b) the distance of the exemplars (data points) from the queries.

---

**Answer**

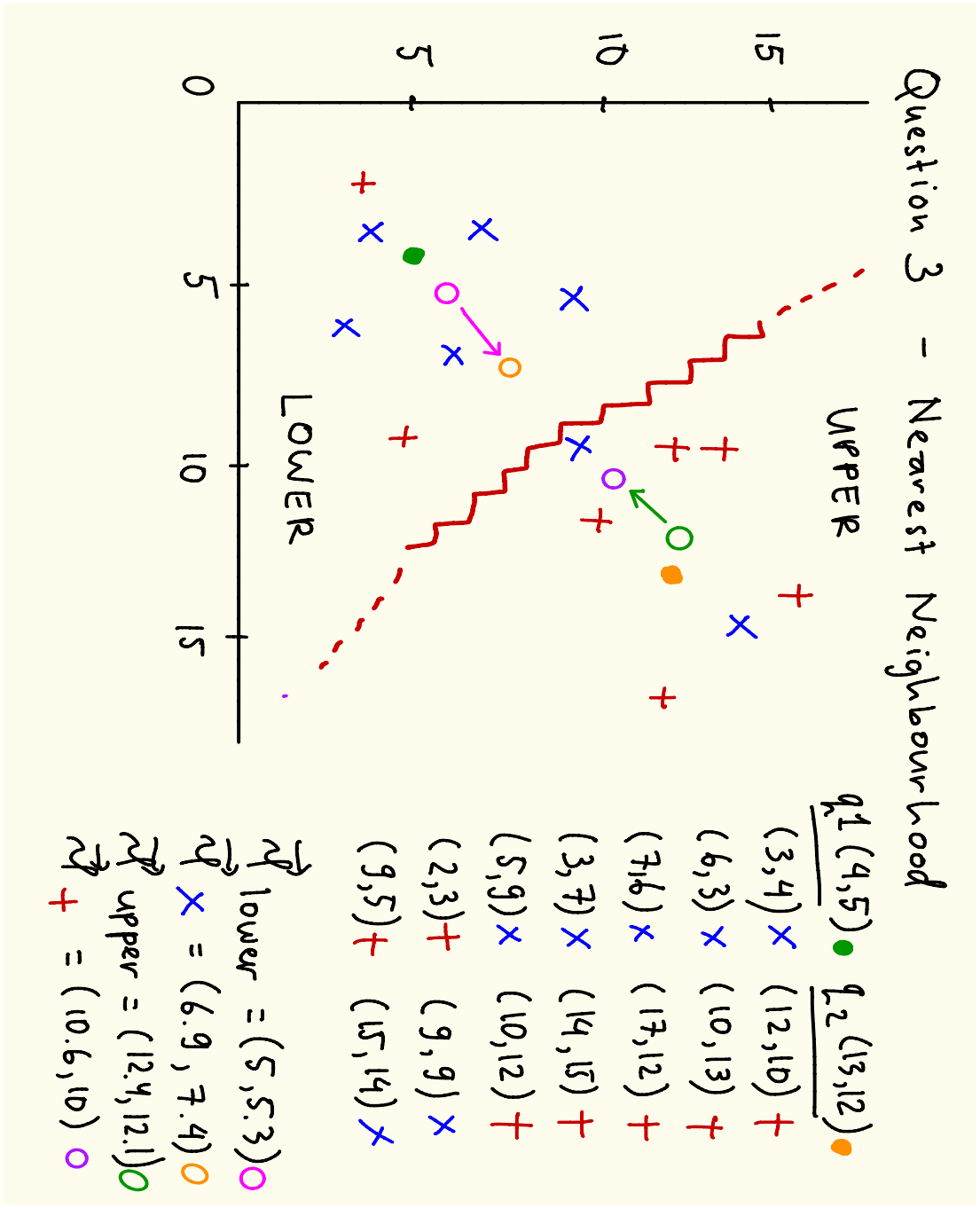



---

**Question 3** Outline how the idea of computing the arithmetic mean $\mu$ of a set of labelled data points (positive or negative) in some Euclidean space can be used to construct a linear classifier. [HINT: consider the basic linear classifier referred to in the lecture notes on slides 15–19.]

---

**Answer**




**Question 4** (**Challenge**) Complete the proof that the arithmetic mean minimises squared Euclidean distance (see slide 15 of the lecture notes). Find the minimum by taking the gradient (vector

of partial derivatives) and setting to zero.

---

**Answer**

TBA. For now, this is left as an exercise.

---

## Naive Bayes classifier

**Question 5** Consider the example application of Bayes Theorem on slides 48–52 in the lecture notes. Suppose the a second laboratory test is ordered for the same patient, and this test also returns a positive result. What are the posterior probabilities of *cancer* and $\neg$*cancer* following these two tests? Note: you can assume that the two tests are independent.

---

**Answer**

Assume: two tests are independent.

Priors $P(\text{cancer}) = 0.008$ $P(\neg\text{cancer}) = 0.992$

Likelihoods $P(\oplus|\text{cancer}) = 0.98$ $P(\ominus|\text{cancer}) = 0.02$ Given

$P(\oplus|\neg\text{cancer}) = 0.03$ $P(\ominus|\neg\text{cancer}) = 0.97$

First test = $\oplus$ $\left[\arg\max_h P(h|D) = P(D|h)P(h)\right]$

$P(\oplus|\text{cancer})\,P(\text{cancer}) = 0.98 \times 0.008 = 0.0078$

$P(\oplus|\neg\text{cancer})\,P(\neg\text{cancer}) = 0.03 \times 0.992 = 0.0298$

$\Rightarrow$ MAP = $\neg$cancer

Second test = $\oplus$ $D = \{\overset{d_1}{\oplus}, \overset{d_2}{\oplus}\}$

$P(d_1|h)\,P(d_2|h)\,P(h)$ // By independence of data!

$P(\oplus|c)\,P(\oplus|c)\,P(c) = 0.98 \times 0.98 \times 0.008 = 0.0077$

$P(\oplus|\neg c)\,P(\oplus|\neg c)\,P(\neg c) = 0.03 \times 0.03 \times 0.992 = 0.0009$

$\Rightarrow$ MAP = cancer

N.B. how quickly probabilities get v. small $\Rightarrow$ log-space calculation

**Question 6** Work through the example of applying Naive Bayes to text on slides 109–118. Be sure you are clear on the difference between the multinomial and multivariate Bernoulli models.

**Answer**
See notes on following pages. First up are two pages with the key steps. The rest is some older handwritten notes with some extra information, but at the cost of poorer legibility – these are just included for completeness.

Question 6 :

(+)

$e_1$ : b d e b b d e

$e_2$ : b c e b b d d e c c

$e_3$ : a d a d e a e e

$e_4$ : b a d b e d a b

(−)

$e_5$ : a b a b a b a e d

$e_6$ : a c a c a c a e d

$e_7$ : e a e d a e a

$e_8$ : d e d e d

MULTINOMIAL
Count Vector

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ | 0 | 3 | 0 | + |
| $e_2$ | 0 | 3 | 3 | + |
| $e_3$ | 3 | 0 | 0 | + |
| $e_4$ | 2 | 3 | 0 | + |
| $e_5$ | 4 | 3 | 0 | − |
| $e_6$ | 4 | 0 | 3 | − |
| $e_7$ | 3 | 0 | 0 | − |
| $e_8$ | 0 | 0 | 0 | − |

MULTIVARIATE BERNOULLI
Bit Vector

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ | 0 | 1 | 0 | + |
| $e_2$ | 0 | 1 | 1 | + |
| $e_3$ | 1 | 0 | 0 | + |
| $e_4$ | 1 | 1 | 0 | + |
| $e_5$ | 1 | 1 | 0 | − |
| $e_6$ | 1 | 0 | 1 | − |
| $e_7$ | 1 | 0 | 0 | − |
| $e_8$ | 0 | 0 | 0 | − |

# Smoothing

MULTINOMIAL
Count Vector

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ | 0 | 3 | 0 | + |
| $e_2$ | 0 | 3 | 3 | + |
| $e_3$ | 3 | 0 | 0 | + |
| $e_4$ | 2 | 3 | 0 | + |
| $v_1$ | 1 | 0 | 0 | + |
| $v_2$ | 0 | 1 | 0 | + |
| $v_3$ | 0 | 0 | 1 | + |
| $e_5$ | 4 | 3 | 0 | − |
| $e_6$ | 4 | 0 | 3 | − |
| $e_7$ | 3 | 0 | 0 | − |
| $e_8$ | 0 | 0 | 0 | − |
| $v_1$ | 1 | 0 | 0 | − |
| $v_2$ | 0 | 1 | 0 | − |
| $v_3$ | 0 | 0 | 1 | − |

MULTIVARIATE BERNOULLI
Bit Vector

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ | 0 | 1 | 0 | + |
| $e_2$ | 0 | 1 | 1 | + |
| $e_3$ | 1 | 0 | 0 | + |
| $e_4$ | 1 | 1 | 0 | + |
| $v_1$ | 1 | 1 | 1 | + |
| $v_2$ | 0 | 0 | 0 | + |
| $e_5$ | 1 | 1 | 0 | − |
| $e_6$ | 1 | 0 | 1 | − |
| $e_7$ | 1 | 0 | 0 | − |
| $e_8$ | 0 | 0 | 0 | − |
| $v_1$ | 1 | 1 | 1 | − |
| $v_2$ | 0 | 0 | 0 | − |

**Bernoulli** "event with two possible outcomes"

$\theta$ prob$^y$ success

mean $\mathbb{E}[X] = \theta$ variance $\mathbb{E}[(X - \mathbb{E}[X])^2] = \theta(1-\theta)$

**Binomial** "number of successes $s$ in $n$ iid trials" (prob$^y$ $\theta$)

$$p(s, n) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

mean $\mathbb{E}[s] = n\theta$ variance $\mathbb{E}[(s - \mathbb{E}[s])^2] = n\theta(1-\theta)$

**Categorical** "event with > 2 possible outcomes"

aka Generalized Bernoulli

or Discrete

$\vec{\theta}$ prob$^y$ of $k$ possible outcomes, $\sum_{i=1}^{k} \theta_i = 1$

**Multinomial** "number of $k$-categorical outcomes in $n$ iid trials" (prob $\vec{\theta}$)

$\vec{X} = (X_1, X_2, \ldots, X_k)$ is a $k$-vector of counts

$$p\left(\vec{X} = (x_1, x_2, \ldots, x_k)\right) =$$

$$n! \frac{\theta_1^{x_1}}{x_1!} \cdots \frac{\theta_k^{x_k}}{x_k!} \quad \text{with} \sum_{i=1}^{k} x_i = n$$

Dirichlet is conjugate prior

Raw data – text, w/ words (emails, Tweets, etc)

$P_i$ cat. $i$

m-est. $\frac{d + p \cdot m}{n + m}$

| ⊕ | ⊖ |
|---|---|
| $e_1$ : b d e bb de | $e_5$ : a b ab ab a e d |
| $e_2$ : b c e bb dd e cc | $e_6$ : a c ac ac a e d |
| $e_3$ : a d a d e a ee | $e_7$ : e a e d a e a |
| $e_4$ : b a d b e d ab | $e_8$ : d e d e d |

"pseudo-counts"

* Laplace $\frac{d}{n}\left(\frac{\#\text{ occ. doc} + 1}{\#\text{ doc} + 2}\right)$

**Counts**

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ | 0 | 3 | 0 | + |
| $e_2$ | 0 | 3 | 3 | + |
| $e_3$ | 3 | 0 | 0 | + |
| $e_4$ | 2 | 3 | 0 | + |
| $e_5$ | 4 | 3 | 0 | − |
| $e_6$ | 4 | 0 | 3 | − |
| $e_7$ | 3 | 0 | 0 | − |
| $e_8$ | 0 | 0 | 0 | − |

Sum (5 4 3) ⊕

(11 3 3) ⊖

Smooth (add 1 word) (6 10 4) ⊕

$\left(\frac{12}{20}\ \frac{4}{20}\ \frac{4}{20}\right)$ ⊖

(0.3 0.5 0.2) ⊕

(0.6 0.2 0.2) ⊖

**Bit vec**

| | a | b | c | class |
|---|---|---|---|---|
| $e_1$ : | 0 | 1 | 0 | + |
| $e_2$ : | 0 | 1 | 1 | + |
| $e_3$ : | 1 | 0 | 0 | + |
| $e_4$ : | 1 | 1 | 0 | + |
| $e_5$ : | 1 | 1 | 0 | − |
| $e_6$ : | 1 | 0 | 1 | − |
| $e_7$ : | 1 | 0 | 0 | − |
| $e_8$ : | 0 | 0 | 0 | − |

Sum (2 3 1) ⊕

(3 1 1) ⊖

Smooth Laplace * [2 pseudo-documents: (1 1 1 ..) & (0 0 0 ..)]

$\left(\frac{3}{6}\ \frac{4}{6}\ \frac{2}{6}\right)$ ⊕

$\left(\frac{4}{6}\ \frac{2}{6}\ \frac{2}{6}\right)$ ⊖

(0.5 0.67 0.33) ⊕

(0.67 0.33 0.33) ⊖

DATA: multivariate Bernoulli $(X_1, X_2, X_3 \ldots)$ $\theta_1\ \theta_2\ \theta_3 \ldots$ multinomial (count vector) $(X_1\ X_2\ X_3 \ldots)$

## Prediction (using multi-variate Bernoulli)

Model parms.

| a | b | c | class |
|---|---|---|---|
| 0.5 | 0.67 | 0.33 | ⊕ |
| 0.67 | 0.33 | 0.33 | ⊖ |

New instance $\vec{x}$:

| a | b | c |
|---|---|---|
| 1 | 1 | 0 |

$h_{ML}$ $\quad \underset{\text{class} \in \{\oplus, \ominus\}}{\text{argmax}} \; P(\vec{x} \mid \text{class})$

$$P(\vec{x} \mid \oplus) = 0.5 * 0.67 * (1 - 0.33) = 0.222$$
$$P(\vec{x} \mid \ominus) = 0.67 * 0.33 * (1 - 0.33) = 0.148$$

↪ predict ⊕

$$LR = \frac{P(\vec{x} \mid \oplus)}{P(\vec{x} \mid \ominus)} = \frac{0.5}{0.67} * \frac{0.67}{0.33} * \frac{1 - 0.33}{1 - 0.33} = 3/2 > 1$$

$h_{MAP}$ predict ⊕ if prior odds $\frac{P(\oplus)}{P(\ominus)} > 2/3$ !