

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)

$$(x_2, y_2)$$

$$(x_1, y_1)$$

$$\sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$$

$$\sqrt{\sum_{i=1}^n \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right)^2}$$

$$Dis_{Euc} (\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

$$Dis_{MINKOWSKI} (\vec{x}, \vec{y}) = \left(\sum_{i=1}^n (x_i - y_i)^p \right)^{\frac{1}{p}}$$

e.g. $p = 1$ MANHATTAN

$p = 2$ EUCLIDEAN

$p = \infty$ "largest component"

$$p\text{-norm} \quad \|\vec{z}\|_p = \left(\sum_i |z_i|^p \right)^{\frac{1}{p}}$$

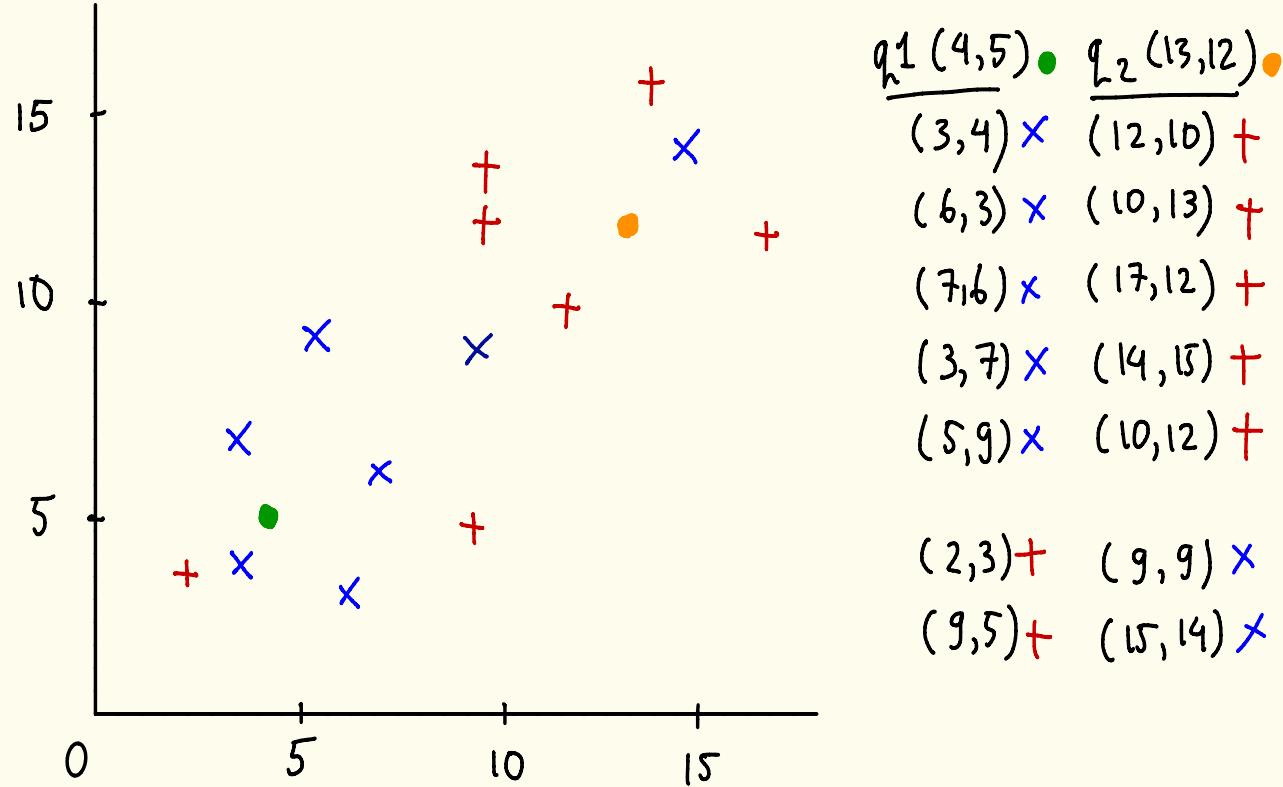
If $\vec{z} = \vec{x} - \vec{y}$ Then $Dis_{Euc} = p\text{-norm}$.

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

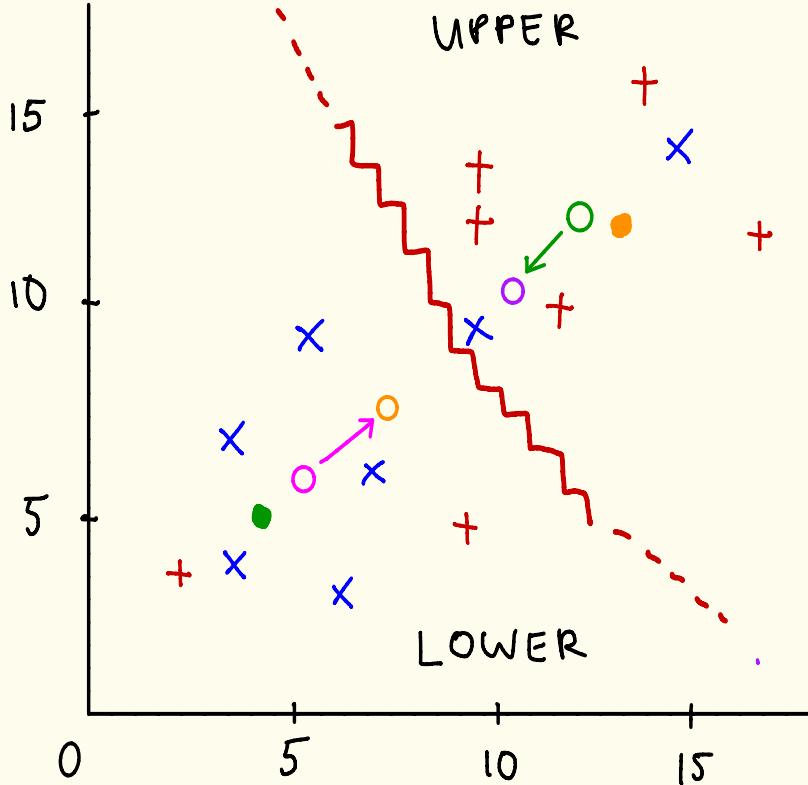
Question 2 - Nearest Neighbourhood



Question 3 Outline how the idea of computing the arithmetic mean μ 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 3 - Nearest Neighbourhood



$$\frac{q_1(4,5)}{(3,4)} \times \frac{q_2(13,12)}{(12,10)} +$$

$$(6,3) \times (10,13) +$$

$$(7,6) \times (17,12) +$$

$$(3,7) \times (14,15) +$$

$$(5,9) \times (10,12) +$$

$$(2,3) + (9,9) \times$$

$$(9,5) + (15,14) \times$$

$$\begin{matrix} \nearrow & \searrow \\ \text{lower} & = (5, 5.3) \end{matrix} \circ$$

$$\begin{matrix} \nearrow & \searrow \\ \text{x} & = (6.9, 7.4) \end{matrix} \circ$$

$$\begin{matrix} \nearrow & \searrow \\ \text{upper} & = (12.4, 12.1) \end{matrix} \circ$$

$$\begin{matrix} \nearrow & \searrow \\ + & = (10.6, 10) \end{matrix} \circ$$

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 .

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

$$\text{Likelihoods} \quad P(+ | \text{cancer}) = 0.98 \quad P(- | \text{cancer}) = 0.02 \quad \text{Given}$$

$$P(+ | \neg \text{cancer}) = 0.03 \quad P(- | \neg \text{cancer}) = 0.97$$

$$\text{First test} = + \quad [\arg \max_h P(h | D) = P(D|h)P(h)]$$

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

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

$$\text{Second test} = + \quad D = \{+, +\} \quad \Rightarrow \underline{\text{MAP}} = \neg \text{cancer}$$

$$P(d_1 | h) P(d_2 | h) P(h) \quad // \text{By independence of data !}$$

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

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

$$\Rightarrow \underline{\text{MAP}} = \text{cancer}$$

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 :

\oplus

e₁ : b d e b b d e

e₂ : b c e b b d d e c c

e₃ : a d a d e a e e

e₄ : b a d b e d a b

\ominus

e₅ : a b a b a b a e d

e₆ : a c a c a c a c a e d

e₇ : e a e d a e a

e₈ : d e 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"

θ prob "success"

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

Binomial

"number of successes s in n trials" ($\text{prob } \theta$)

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

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

Categorical

"event with > 2 possible outcomes"

aka
Generalized
Bernoulli

or
Discrete

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

Multinomial

"number of k -categorical outcomes in n iid trials" ($\text{prod } \vec{\theta}$)

$\vec{X} = (x_1, x_2, \dots, x_k)$ is a k -vector of counts

$$P(\vec{X} = (x_1, x_2, \dots, x_k)) =$$

$$\frac{n!}{x_1! x_2! \dots x_k!} \frac{\theta_1^{x_1}}{x_1!} \dots \frac{\theta_k^{x_k}}{x_k!} \quad \text{with } \sum_{i=1}^k x_i = n$$

Dirichlet
is
conjugate
prior

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

⊕
 $e_1: b \ a \ e \ bb \ de$
 $e_2: b \ c \ e \ bb \ dd \ e \ cc$
 $e_3: a \ d \ a \ d \ e \ a \ ee$
 $e_4: b \ a \ d \ b \ ed \ ab$

$$\theta_i = \frac{d + p_i m}{n + m}$$

$$\theta_i = \frac{d + \text{pseudo-counts}}{n + \text{pseudo-counts}}$$

$$\theta_i = \frac{d(\# \text{occ. doc} + 1)}{n(\# \text{doc} + 2)}$$

	Counts			class
	a	b	c	
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	-

	Bit vec			class
	a	b	c	
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 $(5 \ 9 \ 3) \oplus$
 $(11 \ 3 \ 3) \ominus$

Sum $(2 \ 3 \ 1) \oplus$
 $(3 \ 1 \ 1) \ominus$

Smooth
 (add 1 word) $(6 \ 10 \ 4) \oplus$
 $(\frac{12}{20} \ \frac{4}{20} \ \frac{4}{20}) \ominus$
 $(0.3 \ 0.5 \ 0.2) \oplus$
 $(0.6 \ 0.2 \ 0.2) \ominus$

Smooth Laplace * [2 pseudo-documents: (1111...) & (0000...)]
 $(\frac{3}{6} \ \frac{4}{6} \ \frac{2}{6}) \oplus$
 $(\frac{4}{6} \ \frac{2}{6} \ \frac{2}{6}) \ominus$
 $(0.5 \ 0.67 \ 0.33) \oplus$
 $(0.67 \ 0.33 \ 0.33) \ominus$

DATA: multivariate Bernoulli (X_1, X_2, X_3, \dots) multinomial (count vector) (X_1, X_2, X_3, \dots)

Prediction			(using multi-variate Bernoulli)
	a	b	c
Model parms.	(0.5 0.67 0.33) \oplus		
	(0.67 0.33 0.33) \ominus		

New instance: $\vec{x} = \begin{pmatrix} a \\ 1 \\ 1 \\ 0 \end{pmatrix}$

$$\underline{H_{ML}} \quad \underset{\text{class } c \in \{\oplus, \ominus\}}{\operatorname{argmax}} \quad P(\vec{x} | \text{class } c)$$

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

$$P(\vec{x} | \ominus) = 0.67 * 0.33 * (1 - 0.33) = 0.148$$

\hookrightarrow predict \oplus

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

$$\underline{H_{MAP}} \quad \text{predict } \oplus \text{ if prior odds } \frac{P(\oplus)}{P(\ominus)} > 2/3$$