

Part I: Pen and paper

We collected four positive (P) observations,

$$\{x_1 = (A, 0), \quad x_2 = (B, 1), \quad x_3 = (A, 1), \quad x_4 = (A, 0)\}$$

and four negative (N) observations,

$$\{x_5 = (B, 0), \quad x_6 = (B, 0), \quad x_7 = (A, 1), \quad x_8 = (B, 1)\}$$

Consider the problem of classifying observations as positive or negative.

1. Compute the F1-measure of a kNN with $k = 5$ and Hamming distance using a leave-one-out evaluation schema. Show all calculus.

We start by calculating Hamming distance between observations. The Hamming distance is the number of positions at which the corresponding symbols are different.

Since we are working with $k = 5$, we will consider the 5 nearest neighbors of each observation (written in blue).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	-	2	1	0	1	1	1	2
x_2	2	-	1	2	1	1	1	0
x_3	1	1	-	1	2	2	0	1
x_4	0	2	1	-	1	1	1	2
x_5	1	1	2	1	-	0	2	1
x_6	1	1	2	1	0	-	2	1
x_7	1	1	0	1	2	2	-	1
x_8	2	0	1	2	1	1	1	-

Table 1: Hamming distance between observations

Now that we have the Hamming distance between all observations, we must identify if the prediction is correct or not. We will consider the majority class of the 5 nearest neighbors for each observation.

Example: For x_1 , the 5 nearest neighbors are x_3 and x_4 (which are positive), x_5 , x_6 and x_7 (which are negative). The majority class is negative, therefore the prediction is incorrect.

We apply the same logic for the rest of the classes, ending up with the following table:

Observation	True Value	Prediction	Confusion Matrix Terminology
x_1	P	N	FP
x_2	P	N	FP
x_3	P	P	TP
x_4	P	N	FN
x_5	N	P	FP
x_6	N	P	FP
x_7	N	P	FP
x_8	N	N	TN

P - Positive observation; N - Negative observation

TP - True Positive; TN - True Negative; FP - False Positive; FN - False Negative

Table 2: Predictions for each observation

With this table, we can now calculate the Precision, Recall and F1-measure using the following formulas:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Replacing the corresponding values in the formulas, we get:

for Precision (1) and Recall (2):

$$\text{Precision} = \frac{1}{1 + 5} \approx 0.1667 \quad \text{Recall} = \frac{1}{1 + 1} = 0.5$$

F1-measure (3):

$$\text{F1-measure} = 2 \times \frac{0.1667 \times 0.5}{0.1667 + 0.5} \approx 0.25$$

2. **Propose a new metric (distance) that improves the latter's performance (i.e., the F1-measure) by three fold.**

To improve the performance of the F1-measure, we can use a weighted distance metric that takes into account the type of data we are working with. In this case, we have two types of data: numeric and categorical. The formula for the new metric is:

$$d(x_1, x_2) = \alpha \times d_{\text{numeric}}(x_1, x_2) + \beta \times d_{\text{categorical}}(x_1, x_2)$$

where α and β are parameters that reveal the importance of each component, generally $\alpha + \beta = 1$. The distance between two numeric values is calculated using the Euclidean distance, while the distance between two categorical values is calculated using the Hamming distance.

An additional positive observation was acquired, $x_9 = (B, 0)$, and a third variable y_3 was independently monitored, yielding estimates,

$$y_3|P = \{1.1, 0.8, 0.5, 0.9, 0.8\} \quad \text{and} \quad y_3|N = \{1, 0.9, 1.2, 0.9\}$$

3. **Considering the nine training observations, learn a Bayesian classifier assuming: (i) y_1 and y_2 are dependent; (ii) y_1 , y_2 and y_3 variable sets are independent and equally important; and (iii) y_3 is normally distributed. Show all parameters.**

With the nine training observations, we can calculate the parameters for the Bayesian classifier. We will refer to the outcome, which can be positive or negative, as P and N respectively, and 'class' when we mean both.

Priors:

$$P(P) = \frac{5}{9} \quad P(N) = 1 - P(P) = \frac{4}{9}$$

Likelihoods:

$$\begin{aligned} P((y_1 = A, y_2 = 0)|P) &= \frac{2}{5} & P((y_1 = A, y_2 = 1)|P) &= \frac{1}{5} \\ P((y_1 = A, y_2 = 0)|N) &= \frac{0}{4} & P((y_1 = A, y_2 = 1)|N) &= \frac{1}{4} \\ P((y_1 = B, y_2 = 0)|P) &= \frac{1}{5} & P((y_1 = B, y_2 = 1)|P) &= \frac{1}{5} \\ P((y_1 = B, y_2 = 0)|N) &= \frac{2}{4} & P((y_1 = B, y_2 = 1)|N) &= \frac{1}{4} \end{aligned}$$

For the third variable, we know that it is normally distributed. We can calculate the parameters for each class as follows:

$$\mu_P = \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.82$$

$$\sigma_P^2 = \frac{(1.1 - 0.82)^2 + (0.8 - 0.82)^2 + (0.5 - 0.82)^2 + (0.9 - 0.82)^2 + (0.8 - 0.82)^2}{5} \approx 0.21$$

$$\mu_N = \frac{1 + 0.9 + 1.2 + 0.9}{4} = 1.0$$

$$\sigma_N^2 = \frac{(1 - 1)^2 + (0.9 - 1)^2 + (1.2 - 1)^2 + (0.9 - 1)^2}{4} \approx 0.01$$

The likelihoods for numeric variables with normal distribution are calculated using the following formula:

$$P(y_z|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_z - \mu)^2}{2\sigma^2}\right) \quad (4)$$

We must now replace y_3 with each value in the set $\{1.1, 0.8, 0.5, 0.9, 0.8\}$ and $\{1, 0.9, 1.2, 0.9\}$ to calculate the likelihoods for each class. Having the parameters for the normal distribution, we can calculate the likelihoods for y_3 as per (4):

Positive class:

$$\begin{aligned} P(y_3 = 1.1|\mu_P, \sigma_P^2) &= \frac{1}{\sqrt{2\pi \times 0.21}} \exp\left(-\frac{(1.1 - 0.82)^2}{2 \times 0.21}\right) \approx 0.722 \\ P(y_3 = 0.8|\mu_P, \sigma_P^2) &= \frac{1}{\sqrt{2\pi \times 0.21}} \exp\left(-\frac{(0.8 - 0.82)^2}{2 \times 0.21}\right) \approx 0.870 \\ P(y_3 = 0.5|\mu_P, \sigma_P^2) &= \frac{1}{\sqrt{2\pi \times 0.21}} \exp\left(-\frac{(0.5 - 0.82)^2}{2 \times 0.21}\right) \approx 0.682 \\ P(y_3 = 0.9|\mu_P, \sigma_P^2) &= \frac{1}{\sqrt{2\pi \times 0.21}} \exp\left(-\frac{(0.9 - 0.82)^2}{2 \times 0.21}\right) \approx 0.857 \end{aligned}$$

Negative class:

$$\begin{aligned} P(y_3 = 1|\mu_N, \sigma_N^2) &= \frac{1}{\sqrt{2\pi \times 0.01}} \exp\left(-\frac{(1 - 1)^2}{2 \times 0.01}\right) \approx 0.25 \\ P(y_3 = 0.9|\mu_N, \sigma_N^2) &= \frac{1}{\sqrt{2\pi \times 0.01}} \exp\left(-\frac{(0.9 - 1)^2}{2 \times 0.01}\right) \approx \\ P(y_3 = 1.2|\mu_N, \sigma_N^2) &= \frac{1}{\sqrt{2\pi \times 0.01}} \exp\left(-\frac{(1.2 - 1)^2}{2 \times 0.01}\right) \approx 0.25 \end{aligned}$$

These parameters, allow us to calculate $P(y_3|class)$ using the next expression:

$$P(y_3|\mu_{class}, \sigma_{class}^2) = \prod_{i=1}^n P(y_3(i)|\mu_{class}, \sigma_{class}^2)$$

Replacing (3) the values for each class, we get:

$$P(y_3|\mu_P, \sigma_P^2) \approx blabla \quad P(y_3|\mu_N, \sigma_N^2) \approx blabla$$

To estimate $P(P|y_1, y_2, y_3)$, we can use Bayes' theorem:

$$P(\text{class}|y_1, y_2, y_3) = \frac{P(y_1, y_2, y_3|\text{class}) \times P(\text{class})}{P(y_1, y_2, y_3)} \quad (5)$$

Since we know $\{y_1, y_2\}$, $\{y_3\}$ and are independent, we can rewrite $P(y_1, y_2, y_3)$ as $P(y_1, y_2) \cdot P(y_3)$. Rewriting (5) with this, results in:

$$P(\text{class}|y_1, y_2, y_3) = \frac{P(y_1, y_2|\text{class}) \times P(y_3|\text{class}) \times P(\text{class})}{P(y_1, y_2) \times P(y_3)} \quad (6)$$

Now we can calculate $P(P|y_1, y_2, y_3)$ and $P(N|y_1, y_2, y_3)$ using (6):

Positive class:

$$P(P|y_1, y_2, y_3) = \frac{P(y_1, y_2|P) \times P(y_3|P) \times P(P)}{P(y_1, y_2) \times P(y_3)} = \frac{\frac{1}{5} \times \frac{5}{9}}{\frac{1}{5} \times \frac{4}{9}}$$

Negative class:

$$P(N|y_1, y_2, y_3) = \frac{P(y_1, y_2|N) \times P(y_3|N) \times P(N)}{P(y_1, y_2) \times P(y_3)} = \frac{\frac{1}{4} \times \frac{4}{9}}{\frac{1}{4} \times \frac{4}{9}}$$

Consider now three testing observations,

$$\{(A, 1, 0.8), (B, 1, 1), (B, 0, 0.9)\}$$

4. Under a MAP assumption, classify each testing observation showing all your calculus.

MAP (Maximum A Posteriori) is defined as:

$$\begin{aligned} \hat{z} &= \arg \max_{c_i} \{P(c_i|x)\} \\ &= \arg \max_{c_i} \left\{ \frac{P(x|c_i) \times P(c_i)}{P(x)} \right\} \\ &= \arg \max_{c_i} \{P(x|c_i) \times P(c_i)\} \end{aligned} \quad (7)$$

At last, consider only the following sentences and their respective connotations,

$$\{("Amazing run", P), ("I like it", P), ("To tired", N), ("Bad run", N)\}$$

5. Using a naïve Bayes under a ML assumption, classify the new sentence "I like to run". For the likelihoods calculation consider the following formula,

$$P(T_i|c) = \frac{freq(t_i) + 1}{N_c + V}$$

where t_i represents a certain term i , V the number of unique terms in the vocabulary, and N_c the total number of terms in class c . Show all calculus.

Part II: Programming