Applied Machine Learning: Tutorial Number 3 Answers

School of Informatics, University of Edinburgh
Instructors: Oisin Mac Aodha and Siddharth N.

September 2022

1. In this question your task is to evaluate the performance of two *different* probabilistic binary classifiers, g_1 and g_2. Each classifier takes a vector of features $\boldsymbol{x}_i$ as input and estimates the probability that $\boldsymbol{x}_i$ is from the positive class i.e. $p(y_i = 1 \mid \boldsymbol{x}_i) = g_k(\boldsymbol{x}_i); k = 1, 2$. In this example, you do *not* have direct access to the classifiers, but instead only have their predictions for a set of six validation examples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_6\}$, along with the corresponding ground truth class labels $\{y_1, y_2, ..., y_6\}$.

i	1	2	3	4	5	6
y_i	0	1	1	0	0	1
$g_1(\boldsymbol{x}_i)$	0.1	0.9	0.7	0.3	0.2	0.2
$g_2(\boldsymbol{x}_i)$	0.2	0.2	0.6	0.7	0.6	0.3

i. Report the F-measure (i.e. F-score) for the two classifiers using their predictions above with a threshold of 0.5. Which of the two classifiers performs better on this validation set?

ii. For the second classifier only (i.e. g_2), report the false positive rate and true positive rate corresponding to the thresholds 0.0, 0.33, 0.66, and 1.0.

iii. Sketch the resulting ROC plot for g_2. You only need to show the points on the curve for the thresholds 0.0, 0.33, 0.66, and 1.0. Do you have any observations from looking at your plot?

Solution:

i. The F-measure is the harmonic mean of precision (Pr) and recall (Re):
$F1 = 2\frac{\text{Pr}\cdot\text{Re}}{\text{Pr}+\text{Re}}$, where $\text{Pr} = \frac{TP}{TP+FP}$ and $\text{Re} = \frac{TP}{TP+FN}$.

i	1	2	3	4	5	6
y_i	0	1	1	0	0	1
$g_1(\boldsymbol{x}_i)$	0.1	0.9	0.7	0.3	0.2	0.2
$g_2(\boldsymbol{x}_i)$	0.2	0.2	0.6	0.7	0.6	0.3
$g_1(\boldsymbol{x}_i) > 0.5$	0	1	1	0	0	0
$g_2(\boldsymbol{x}_i) > 0.5$	0	0	1	1	1	0

F-measure for g_1
$\text{Pr} = \frac{2}{2+0}$ and $\text{Re} = \frac{2}{2+1}$
$F1 = 2\frac{1.0\cdot 0.66}{1.0+0.66} \sim 4/5$

F-measure for g_2
$\text{Pr} = \frac{1}{1+2}$ and $\text{Re} = \frac{1}{1+2}$
$F1 = 2\frac{1.0\cdot 0.66}{1.0+0.66} \sim 1/3$
This indicates that g_1 is a better classifier for this validation set using this measure.

ii. Here we simply have to threshold each of the outputs for g_2 using the four different threshold values provided. We use the following expressions for false positive rate (FPR) and true positive rate (TPR). FPR $= \frac{FP}{FP+TN}$
TPR $=$ Re $= \frac{TP}{TP+FN}$

thresh	FPR g_2	TPR g_2
0.0	1	1
0.33	2/3	1/3
0.66	1/3	0
1.0	0	0

iii. The g_2 classifier performs worse than random guessing i.e. it is below the diagonal line. In practice, if we inverted the output of the model (i.e. $p(y_i = 1 \mid \boldsymbol{x}_i) = 1.0 - g_2(\boldsymbol{x}_i)$), we would obtain better performance.

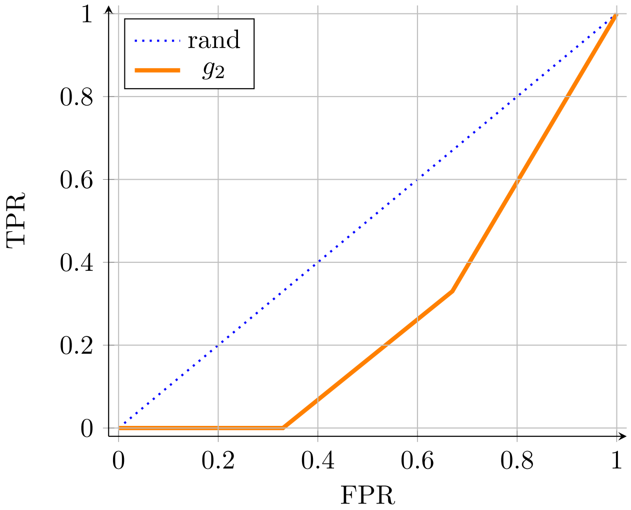

Figure 1: ROC curve for the binary classifier g_2.

2. The Peak Expiratory Flow Rate (PEFR) is a person's maximum speed of expiration, as measured with a peak flow meter, a small, hand-held device used to monitor a person's ability to breathe out air. It measures the airflow through the bronchi and thus the degree of obstruction in the airways.

A medical practitioner at the respiratory illness unit of a leading hospital is unhappy about having people physically blow into a flow meter due to hygiene issues and constraints on the reuse of measurement devices. They find that using scans of patients' chests and throats, they can reliably extract features that identify how wide the bronchi are, any obstructions, and how large lung capacity is. They decide to construct a regression model which they call PERP (short for Peak Expiratory Flow Predictor), that predicts PEFR using these extracted features.

Having trained PERP on some available hospital data, they want to see if their model is actually any good in practice, and so they setup an experiment with four patients that are visiting the hospital that day. They use a flow meter to measure PEFR for each patient directly, then use PERP to try and predict the same using their scans, obtaining the following data:

Patient	Flow Meter	PERP
1	15.22	17.96
2	13.74	16.17
3	22.14	23.81
4	21.66	27.28

i. Are PERP's predictions of PEFR significantly different from the flow meter's predictions? Conduct a t-test to answer this at the $\alpha = 0.05$ significance level. Use the table at the end of this document to identify the critical value c. State the null and alternative hypothesis clearly, and work through the steps required to evaluate the t-test in full.

ii. Suppose that it was found that PERP's estimation had an inadvertent error in that the intercept was estimated incorrectly, and fixing it would mean all of PERP's predictions would decrease by 0.5. Would that affect your answers from the previous part? Would the t-test's final outcome change?

Solution:

Note, we are given $N = 4, \alpha = 0.05$.

i.

$$\begin{aligned} H_0 &: \mu^d = 0 \\ H_1 &: \mu^d \neq 0 \\ c &= 3.182 \qquad \text{(two-tailed)} \end{aligned}$$

Computing the difference in measurements/predictions, we have

$$\begin{aligned} d &= 2.74, 2.43, 1.67, 5.62 \\ \bar{d} &= 3.116 \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2 \\ &= 2.99 \\ s &= 1.73 \\ t &= \frac{\bar{d} - 0}{s/\sqrt{N}} \\ &= 3.599 \end{aligned}$$

$$|t| > c \implies \text{reject } H_0$$

ii. With the correction, the differences all reduce by the exact same amount—0.5.

$$\begin{aligned} d &= 2.24, 1.93, 1.17, 5.12 \\ \bar{d} &= 2.615 \\ s^2 &= 2.99 \\ s &= 1.73 \\ t &= 3.024 \end{aligned}$$

$$|t| \leq c \implies \text{cannot reject } H_0$$

3. Two students are working on a machine-learning approach to spam detection for a large bank. Each student has their *own* set of 100 labeled emails, 90% of which are used for training and 10% for validating the model. Student A runs the Naive Bayes algorithm and reports 80% accuracy on her validation set. Student B experiments with over 100 different learning algorithms, training each one on her training set, and recording the accuracy on the validation set. Her best model achieves 90% accuracy.

i. Whose algorithm would you pick for protecting a corporate network from spam and why?

Solution:

The question is ill-posed: each student has **their own** set of emails. Accuracy figures cannot be directly compared against two different testing sets. If we assume the two students are using the same dataset, and the same training / validation split, the following argument could be made.

There are only 100 labelled emails, so the validation set has only 10 emails in it. It is true that the *expected* performance on the validation set equals the generalisation error, but note that with such a small validation set one would expect quite large variance relative to the true generalisation error.

When student B selects from 100 learning algorithms, she will very likely select on the basis of idiosyncracies in the validation set rather than the true generalisation error.