Calidad modelos regresión

cuadrado de los residuos (los errores que

cometemos al predecir el target de cada

ejemplo) con la dispersión de los targets

R^2=1 sólo si todos los residuos son 0. Por

tanto, **cuanto más cercano a 1, mejor es**

nuestro modelo.

Positive (P) True positive (TP) False negative (FN

Negative (N) False positive (FP) True negative (TN)

• **Precision**: Precision=1 ~ todos los ejemplos estimados positivos lo eran.

F1-score: Media entre precision y recall
Accuracy: porcentaje aciertos

Curva ROC y AUROC

mayor es AUROC, mejor es el modelo.

model = LogisticRegression()

Cálculo de las probabilidades

Cálculo de la curva ROC y AUC

Visualización de la curva ROC

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

y_score = model.predict_proba(testX)[:, 1]

fpr, tpr, thresholds = roc_curve(testY, y_score)

model.fit(trainX, trainY)

roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr)

Se hace entrenamiento con regresión logística:

from sklearn.metrics import roc_curve, auc, confusion_matrix

Entrenamiento del modelo de regresión logística

• Recall: Recall=1 ~todos los ejemplos positivos han sido estimados correctamente.

Positive (P)

True positive (TP),
hit

(FN),
type II error, miss,
underestimation

Sensitivity (SEN),
probability of detection, hit rate,
power $= \frac{TP}{P} = 1 - FNR$ False fregative rate (FNR),
miss rate $= \frac{FN}{P} = 1 - TPR$

Negative (N) type I error, false alarm, (TN), probability of false alarm, fall-out specificity (SPC), selectivity

overestimation correct rejection $= \frac{FP}{N} = 1 - TNR$ $= \frac{TN}{N} = 1 - FPR$

 $\begin{array}{c} \text{Prevalence} \\ = \frac{P}{P+N} \end{array} = \begin{array}{c} \text{Positive predictive value (PPV),} \\ = \frac{TP}{PP} = 1 - FDR \end{array} = \begin{array}{c} \text{False omission} \\ = \frac{FN}{PN} = 1 - NPV \end{array} = \begin{array}{c} \text{Positive likelihood ratio (LR+)} \\ = \frac{TPR}{FPR} \end{array} = \begin{array}{c} \text{Negative likelihood ratio} \\ (LR-) \\ = \frac{FNR}{TNR} \end{array}$

 $\begin{array}{l} \text{Balanced} \\ \text{accuracy (BA)} \\ = \frac{\text{TPR} + \text{TNR}}{2} \end{array} \\ = \frac{2 \text{ PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} \\ = \frac{2 \text{ TP}}{2 \text{ TP} + \text{FP} + \text{FN}} \\ \end{array} \\ \begin{array}{l} \text{Fowlkes-Mallows} \\ \text{index (FM)} \\ = \sqrt{\text{PPV} \times \text{TPR}} \\ \end{array} \\ \begin{array}{l} \text{Matthews correlation coefficient} \\ \text{(MCC)} \\ = \sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} \\ - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}} \\ \end{array} \\ \begin{array}{l} \text{Jaccard index} \\ = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \\ \end{array} \\ \end{array}$

Accuracy (ACC) False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$ False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$ False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$ Markedness (MK), deltaP (ΔP) $= \frac{TN}{P} = 1 - FOP$

Dado un modelo que estima la probabilidad de p(y=1|x), tal que yhat=1 si p(y=1|x) > U;

la curva ROC es el trazado de pares (FPR Y TPR) cuando hacemos variar el umbral *U*.

La curva ROC es diferente a la mtriz de confusión porque se calcula para todo el rango

de posibles umbrales U, mientras que la matriz de confusión ya asume un *U* fijado.

Midiendo el área bajo la curva ROC (AUROC) podemos comparar modelos: cuanto

False positive (FP), True negative False positive rate (FPR), True negative rate (TNR),

NOTA: También se puede utilizar la propia función de pérdida, el "problema" es que NO

devuelve un valor acotado, dependerá de los datos. R^2 siempre tiene valor máximo 1.

Calidad modelos de clasificación binaria:

La métrica típica en regresión es el valor:

sklearn.metrics.r2_score

→ Matriz de confusión

12- Evaluación de

modelos



