Foundations of Data Science, Fall 2020

11a. Multiclass Classification, Measuring Performance

Prof. Dan Olteanu

DaST Data•(Systems+Theory)

University of Zurich[UZH]

Nov 6, 2020

https://lms.uzh.ch/url/RepositoryEntry/16830890400

https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09

Multiclass Classification

Number of classes: $C > 2$

In practice, one of the following approaches is common

One-vs-One:

- Train $\binom{C}{2}$ different classifiers for all pairs of classes
- For new input: Choose the most common classification for the classifiers

One-vs-Rest:

- Train C different classifiers, one class vs the rest $C - 1$
- Typical for classifiers that yield class membership probability or scores
- For new input $\mathbf{x}_{new}$: Pick the class with the largest probability or score

  Break ties by value of $\mathbf{w} \cdot \mathbf{x}_{new} + w_0$

Multiclass Classification

One-vs-One

- Training $K(K-1)/2$ classifiers
- Each training procedure only uses on average $\frac{2}{K}$ of the training data
- "Natural" learning problems are

One-vs-Rest

- Training only K classifiers
- Each training procedure uses the entire training data
- Less "natural" learning problems (#negative points $\gg$ #positive points)

A more efficient method: **Reducing Multiclass to Binary.** *E. Allwein, R. Schapire, Y. Singer.* ICML'00 best paper award.

- Divide classes into pairs of disjoint subsets
- Train a binary classifier to separate the subsets of each pair
- Use an error correcting approach to determine the class label

Measuring Performance

Regression: Same loss function applied to test data as for training

Classification: Number of misclassified data points (classification error)

However, not all mistakes are equally problematic

- Mistakenly blocking a legitimate comment vs failing to mark abuse on online message boards
- Failing to detect medical risk vs inaccurately predicting chance of risk

Classification using logistic regression: $p(y = 1 \mid \mathbf{x}_{new}, \mathbf{w}) = \sigma(\mathbf{w} \cdot \mathbf{x}_{new})$

- We used threshold 0.5 to label a point positive
- If we want very few false positives, we raise the threshold at 0.9: *Predict something as positive only if it were 90% sure*

Decision boundary for generative models: $p(y = 1 \mid \mathbf{x}_{new}, \mathbf{w}) / p(y = 0 \mid \mathbf{x}_{new}, \mathbf{w})$

- We used ratio 1 to treat all errors equally
- Change the ratio if one type of errors is more costly than the other

Measuring Performance for Binary Classification

Confusion Matrix

Prediction	Actual Labels: yes	no
yes	True Positive	False Positive
no	False Negative	True Negative

- True Positive Rate, Sensitivity, Recall: $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$

  TPR = Ratio of true positives to actual positives

- False Positive Rate, Fall-out: $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$

  FPR = Ratio of false positives to actual negatives

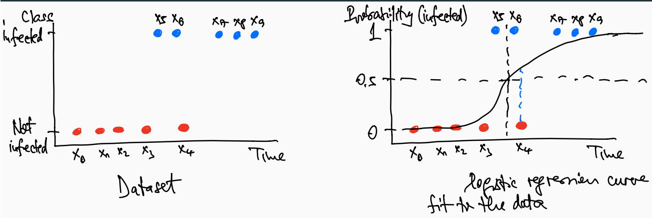
  True negative rate, Specificity = 1 - FPR

- Precision: $\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$

  Precision = Ratio of true positives to predicted positives

- Accuracy = (true positives + true negatives)/(positives + negatives)

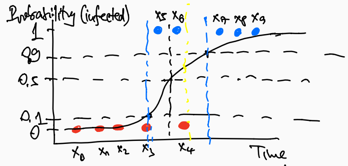
Example: Viral Infections as Function of Time without Mask in Tram



Confusion matrix: for threshold 0.5

Predicted	Actual: infected	Not infected
infected	4	1
Not infected	1	4

Confusion Matrices for Different Decision Boundaries



Threshold 0.1:

Predicted \ Actual	infected	Not infected
infected	5	2
Not infected	0	3

FP↑

FN↓

TN↓

Threshold 0.9:

Predicted \ Actual	infected	Not infected
infected	3	0
Not infected	2	5

$\boxed{FP = 0}$

Which Decision Boundary is Best? Receiver Operating Characteristic (ROC)

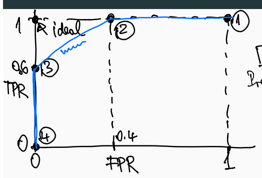
- We only need to try out those thresholds that make a difference
- Instead of analysing many confusion matrices, ROC curves gives an intuitive compact representation of all of them
- ROC space defined by **T**rue **P**ositive **R**ate vs **F**alse **P**ositive **R**ate
- TPR (sensitivity): What proportion of infections were correctly classified?

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- FPR (1-specificity): What proportion of not infected were incorrectly classified?

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

Creating the ROC Curve for several Confusion Matrices



(1) $\boxed{0}$ All infected.

Predicted \ Actual	infected	Not infected
infected	5	5
Not infected	0	0

$TPR = \frac{5}{5+0} = 1.$

$FPR = \frac{5}{5+0} = 1.$

(2) $\boxed{0.1}$

Predicted \ Actual	infected	Not infected
infected	5	2
Not infected	0	3

$TPR = \frac{5}{5+0} = 1.$

$FPR = \frac{2}{2+3} = \frac{2}{5} = 0.4$

(3) $\boxed{0.9}$

Predicted \ Actual	infected	Not infected
infected	3	0
Not infected	2	5

$TPR = \frac{3}{3+2} = \frac{3}{5} = 0.6$

$FPR = \frac{0}{0+5} = 0.$

(4) $\boxed{1}$ NO INF

Predicted \ Actual	infected	Not infected
infected	0	0
Not infected	5	5

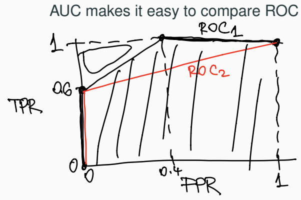
$TPR = \frac{0}{0+5} = 0.$

$FPR = \frac{0}{0+5} = 0.$

(2) > (1).
(3) > (4).
Overall (2) best if we tolerate FP. Otherwise (3).

Area under the ROC Curve (AUC)

AUC makes it easy to compare ROC curves for different classifiers

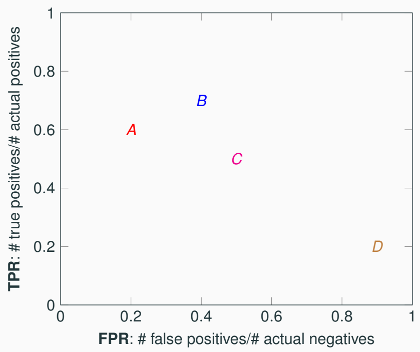


$AUC_1 = 1 - \frac{0.4 \cdot 0.4}{2} = 0.92$

$AUC_2 = 1 - \frac{0.4 \cdot 1}{2} = 0.8$

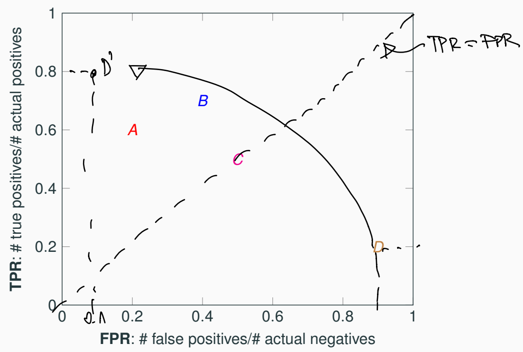
$AUC_1 > AUC_2 \Rightarrow$ classifier with ROC_1 preferred over classifier with ROC_2.

ROC Quiz 1



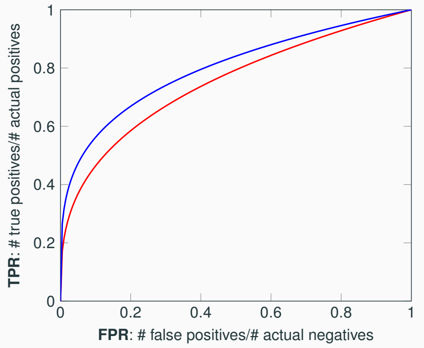
Which of the classifiers *A*, *B*, *C*, or *D* would you consider as most accurate?

ROC Quiz 1



Which of the classifiers *A*, *B*, *C*, or *D* would you consider as most accurate?

ROC Quiz 2



- Plot FPR vs TPR as functions of the threshold t for the decision boundary
- Which curve (red or blue) corresponds to a better trade-off?

Precision-Recall Curves

Another metric beyond ROC curves: Replace FPR with Precision

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision gives the proportion of positive results that were correctly classified

Precision preferred over False Positive Rate in some scenarios

- If there were lots of negative samples
- Precision does not include **T**rue **N**egatives, so not affected by imbalance
- In practice: Studying a rare events, e.g., a rare disease. There are many more samples that do not observe the event than those that observe it.

Measuring Performance for Multiclass Classification

Recall the confusion matrix for binary classification:

	Actual Labels	
Prediction	yes	no
yes	true positive	false positive
no	false negative	true negative

For multi-class classification, we generalise it:

	Actual Labels			
Prediction	1	2	$\cdots$	K
1	N_{11}	N_{12}	$\cdots$	N_{1K}
2	N_{21}	N_{22}	$\cdots$	N_{2K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
K	N_{K1}	N_{K2}	$\cdots$	N_{KK}

$N_{i,j}$: # items of class j in the dataset that were predicted to be of class i

Good classifier: large diagonal entries and small off-diagonal entries