FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

METRICS FOR EVALUATION

FATEMEH AHMADI

Confusion Matrix

Consider a binary classification problem with two classes of P and N:

- **True positives** (*TP*): These refer to the positive tuples that were correctly labeled by the classifier. Let *TP* be the number of true positives.
- **True negatives** (*TN*): These are the negative tuples that were correctly labeled by the classifier. Let *TN* be the number of true negatives.
- **False positives** (*FP*): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class $buys\_computer = no$ for which the classifier predicted $buys\_computer = yes$). Let *FP* be the number of false positives.
- **False negatives** (*FN*): These are the positive tuples that were mislabeled as negative (e.g., tuples of class $buys\_computer = yes$ for which the classifier predicted $buys\_computer = no$). Let *FN* be the number of false negatives.

Confusion Matrix

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P'*	*N'*	*P + N*

Confusion matrix, shown with totals for positive and negative tuples.

Classes	*buys_computer = yes*	*buys_computer = no*	*Total*
buys_computer = yes	**6954**	**46**	7000
buys_computer = no	**412**	**2588**	3000
Total	7366	2634	10,000

Confusion matrix for the classes *buys_computer = yes* and *buys_computer = no*, where an entry in row *i* and column *j* shows the number of tuples of class *i* that were labeled by the classifier as class *j*. Ideally, the nondiagonal entries should be zero or close to zero.

Metrics for Evaluation

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Metrics for Evaluation: Error Rate

We can also speak of the **error rate** or **misclassification rate** of a classifier, M, which is simply $1 - accuracy(M)$, where $accuracy(M)$ is the accuracy of M. This also can be computed as

$$error\ rate = \frac{FP + FN}{P + N}. \quad (8.22)$$

If we were to use the training set (instead of a test set) to estimate the error rate of a model, this quantity is known as the **resubstitution error**. This error estimate is optimistic of the true error rate (and similarly, the corresponding accuracy estimate is optimistic) because the model is not tested on any samples that it has not already seen.

Metrics for Evaluation: Sensitivity and Specificity

The **sensitivity** and **specificity** measures can be used, respectively, for this purpose. Sensitivity is also referred to as the *true positive (recognition) rate* (i.e., the proportion of positive tuples that are correctly identified), while specificity is the *true negative rate* (i.e., the proportion of negative tuples that are correctly identified). These measures are defined as

$$sensitivity = \frac{TP}{P} \tag{8.23}$$

$$specificity = \frac{TN}{N}. \tag{8.24}$$

It can be shown that accuracy is a function of sensitivity and specificity:

$$accuracy = sensitivity \frac{P}{(P+N)} + specificity \frac{N}{(P+N)}. \tag{8.25}$$

Metrics for Evaluation: Precision and Recall

The *precision* and *recall* measures are also widely used in classification. **Precision** can be thought of as a measure of *exactness* (i.e., what percentage of tuples labeled as positive are actually such), whereas **recall** is a measure of *completeness* (what percentage of positive tuples are labeled as such). If recall seems familiar, that's because it is the same as sensitivity (or the *true positive rate*). These measures can be computed as

$$precision = \frac{TP}{TP+FP} \tag{8.26}$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{P}. \tag{8.27}$$

Metrics for Evaluation: F-Measure

In F1, β=1

$$F = \frac{2 \times precision \times recall}{precision + recall} \tag{8.28}$$

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}, \tag{8.29}$$

where β is a non-negative real number. The F measure is the *harmonic mean* of precision and recall (the proof of which is left as an exercise). It gives equal weight to precision and recall. The F_β measure is a weighted measure of precision and recall. It assigns β times as much weight to recall as to precision. Commonly used F_β measures are F_2 (which weights recall twice as much as precision) and $F_{0.5}$ (which weights precision twice as much as recall).

Cross-Validation

In ***k*-fold cross-validation**, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," $D_1, D_2, \ldots, D_k$, each of approximately equal size. Training and testing is performed k times. In iteration i, partition D_i is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets $D_2, \ldots, D_k$ collectively serve as the training set to obtain a first model, which is tested on D_1; the second iteration is trained on subsets $D_1, D_3, \ldots, D_k$ and tested on D_2; and so on. Unlike the holdout and random subsampling methods, here each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data.

Cross-Validation: 10-Fold Cross Validation

	Train_ set	Test_set
D1		√
D2	√	
D3	√	
D4	√	
D5	√	
D6	√	
D7	√	
D8	√	
D9	√	
D10	√	

First Iteration

	Train_ set	Test_set
D1	√	
D2		√
D3	√	
D4	√	
D5	√	
D6	√	
D7	√	
D8	√	
D9	√	
D10	√	

Second Iteration

……

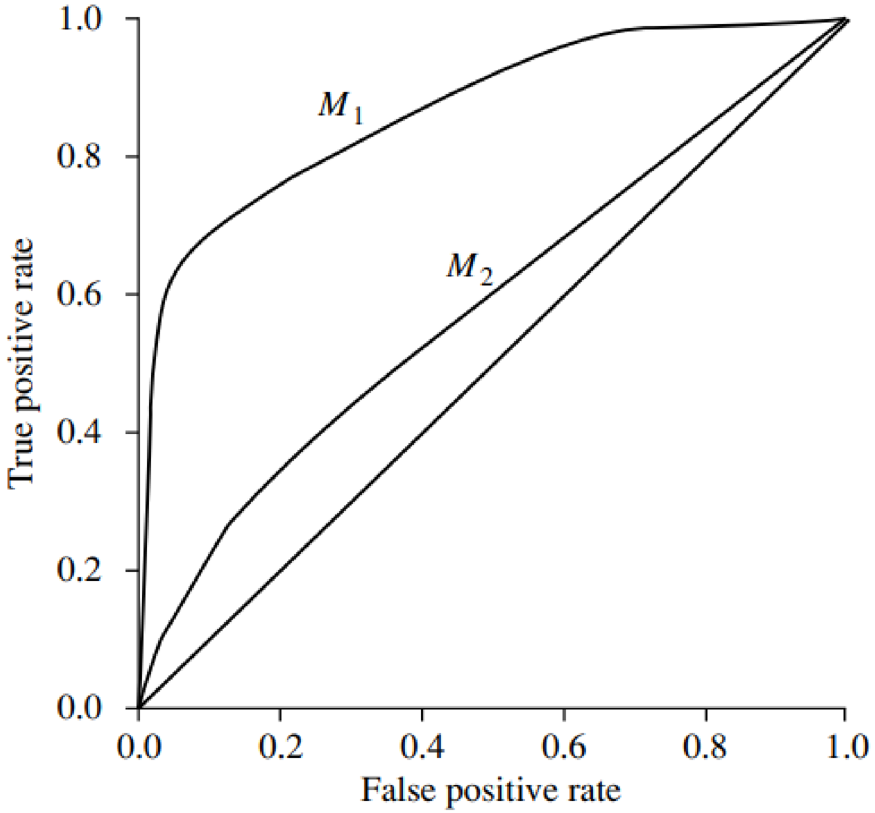
	Train_ set	Test_set
D1	√	
D2	√	
D3	√	
D4	√	
D5	√	
D6	√	
D7	√	
D8	√	
D9	√	
D10		√

10^{th} Iteration

ROC Curve

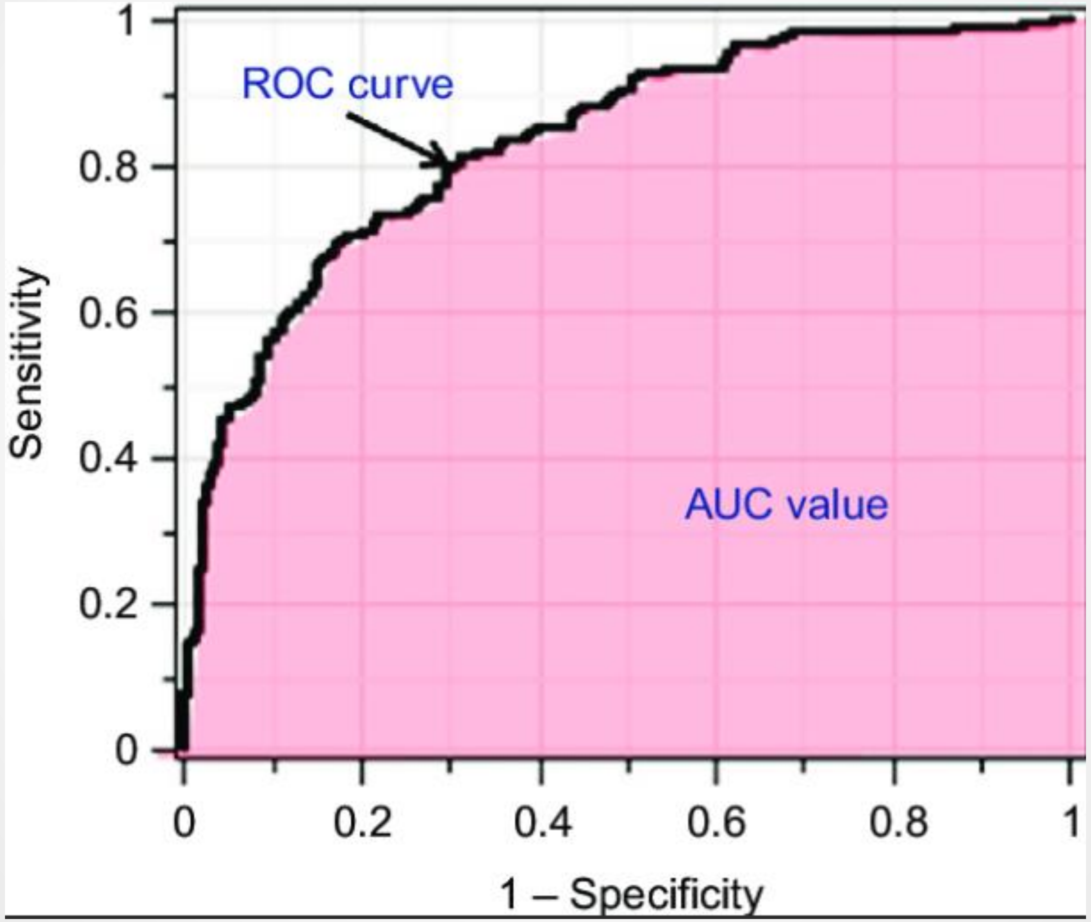
- Next figure shows the ROC curves of two classification models. The diagonal line representing random guessing is also shown. Thus, the closer the ROC curve of a model is to the diagonal line, the less accurate the model.
- If the model is really good, initially we are more likely to encounter true positives as we move down the ranked list. Thus, the curve moves steeply up from zero. Later, as we start to encounter fewer and fewer true positives, and more and more false positives, the curve eases off and becomes more horizontal.
- To assess the accuracy of a model, we can measure the area under the curve. Several software packages are able to perform such calculation. The closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy will have an area of 1.0.

ROC Curve



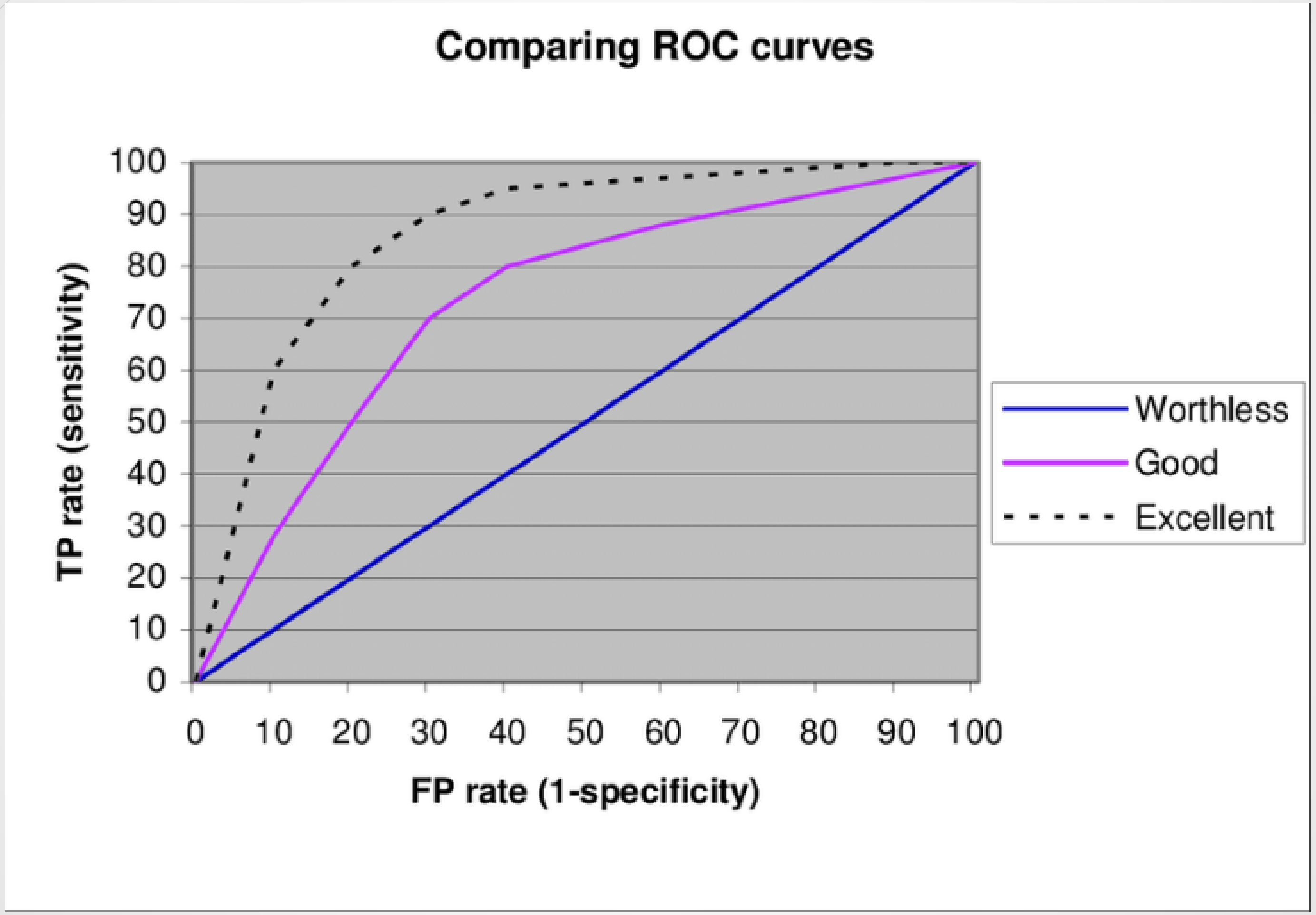
ROC curves of two classification models, M_1 and M_2. The diagonal shows where, for every true positive, we are equally likely to encounter a false positive. The closer an ROC curve is to the diagonal line, the less accurate the model is. Thus, $M1$ is more accurate here.

AUC (Area Under Curve)



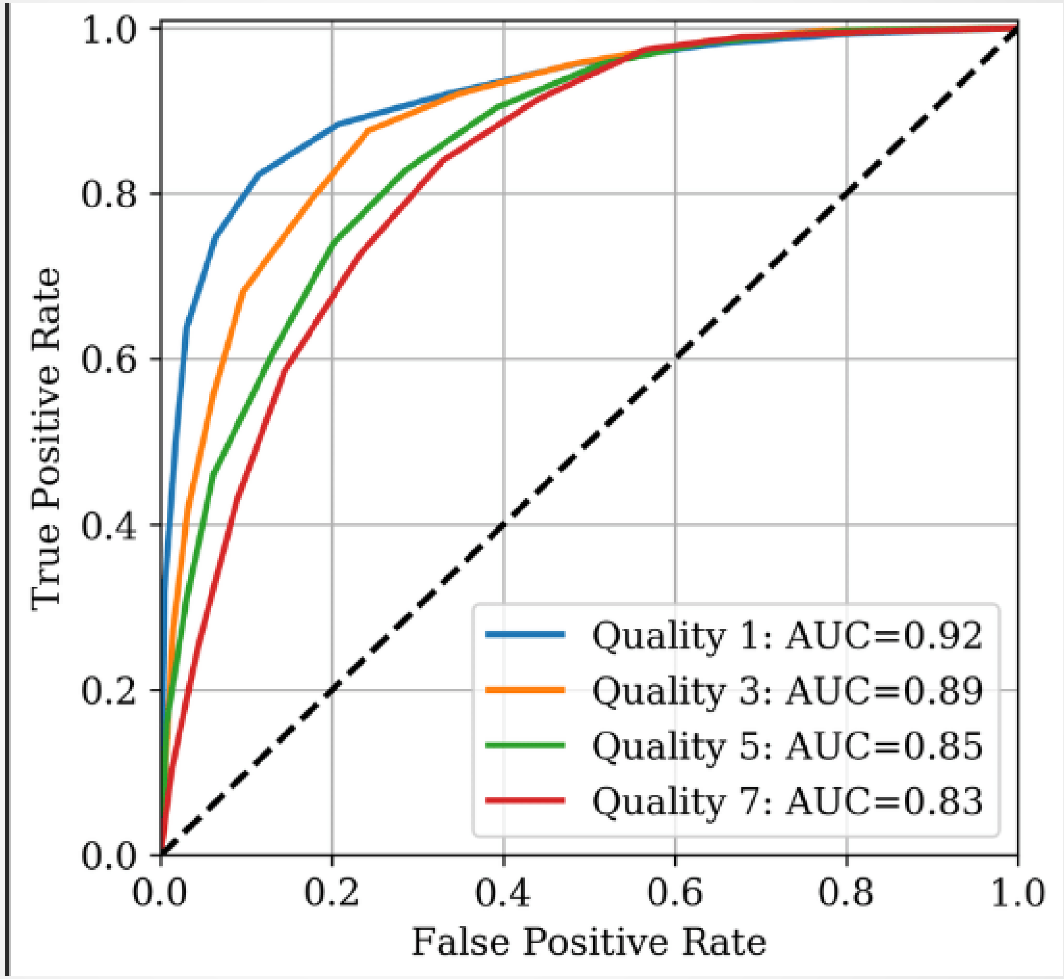
https://www.bing.com/images/search?view=detailV2&ccid=dfOXc6yy&id=F4B79A138400DC8840351A51FD2386CCB94548D2&thid=OIP.dfOXc6yyQ2EGpYQOTn01QQHaGP&mediaurl=https%3A%2F%2Fwww.researchgate.net%2Fprofile%2FBlagoj_Ristevski%2Fpublication%2F275152323%2Ffigure%2Ffig6%2FAS%3A735542471827460%401552378399588%2FThe-ROC-curve-and-AUC-value.png&cdnurl=https%3A%2F%2Fth.bing.com%2Fth%2Fid%2FR.75f39773acb2436106a5840e4e7d3541%3Frik%3D0khFucyGI%252f1RGg%26pid%3DImgRaw%26r%3D0&exph=509&expw=604&q=auc+and+roc+curve&simid=608055554455391020&form=IRPRST&ck=E2F42F586EAA36C8B1F5FC60CC0B3076&selectedindex=0&ajaxhist=0&ajaxserp=0&pivotparams=insightsToken%3Dccid_6NWhxrxi*cp_7F0C5E682621273F635582B48DDC82AE*mid_7A6570856F68EE98F18A49260E72014C71F28DEA*simid_608021989278100931*thid_OIP.6NWhxrxiv8pw1jaBHIZrnAAAAA&vt=0&sim=11&iss=VSI&ajaxhist=0&ajaxserp=0

ROC Curve



https://www.bing.com/images/search?view=detailV2&ccid=eDzQgU69&id=2675D7485123AEA3833E23B576D271D13AD16D96&thid=OIP.eDzQgU692oR7hyP_V7S-8wHaFM&mediaurl=https%3a%2f%2fwww.researchgate.net%2fprofile%2fAnat_Ben-Simon%2fpublication%2f255556204%2ffigure%2fdownload%2ffig1%2fAS%3a669255263600655%401536574297664%2fshows-three-ROC-curves-representing-excellent-good-and-worthless-predictors-The.png&cdnurl=https%3a%2f%2fth.bing.com%2fth%2fid%2fR.783cd0814ebdda847b8723ff57b4bef3%3frik%3dlm3ROtFx0na1Iw%26pid%3dImgRaw%26r%3d0&exph=596&expw=850&q=what+Roc+curve+shows+a+good+model&simid=608053999669550201&FORM=IRPRST&ck=4FDE7B1A568D6E9C61E4F474F27D17DC&selectedIndex=1&ajaxhist=0&ajaxserp=0

ROC Curve



https://www.bing.com/images/search?view=detailV2&ccid=vbyC5Bu%2f&id=A1D7F07350E0F0EC4D90BE82D4CE00BE84FD30D6&thid=OIP.vbyC5Bu_S_v5Ot2PI9h04wHaG7&mediaurl=https%3a%2f%2fwww.researchgate.net%2fprofile%2fNicolo_Bonettini%2fpublication%2f330796849%2ffigure%2fdownload%2ffig3%2fAS%3a726513330696196%401550225684580%2fROC-curves-showing-the-different-possible-working-points-for-each-dataset-pair-according.ppm&cdnurl=https%3a%2f%2fth.bing.com%2fth%2fid%2fR.bdbc82e41bbf4bfbf93add8f97d874e3%3frik%3d1jD9hL4AztSCvg%26pid%3dImgRaw%26r%3d0&exph=796&expw=850&q=what+Roc+curve+shows+a+good+model&simid=608056061246319159&FORM=IRPRST&ck=45FC1BBE5F15E6429FFE0710327B937E&selectedIndex=4&ajaxhist=0&ajaxserp=0

Reference

Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 6.