Evaluating ML Predictions

Expert Testimony

Classification

Classification

- Predicting a categorical variable is **classification**.
- Examples: customer churn, loan default, having a disease, winner-takes all election.
- We cannot use RMSE, correlation or R^2.

Contingency table

Recall **contingency table** of two categorical variables.

	Weather		Total
Pavement	rain	no rain	
wet	8	4	12
dry	0	18	18
Total	8	22	**30**

n_{ij}: number of observations ("cases", "records") when row$=i$ *and* column$=j$

Confusion matrix

The **confusion matrix** is the contingency table of predicted vs actual category.

	Actual		Total
Predicted	does rain	doesn't rain	
will rain	**8**	4	12
won't rain	0	**18**	18
Total	8	22	**30**

Goodness of fit

- **Accuracy** is the fraction of correctly predicted cases $= (8 + 18)/30$.
- But now we can also explore the direction of our error.

	Actual	
Predicted	positive	negative
positive	TP	FP
negative	FN	TN

Good ratios

- **Sensitivity**: probability of positive "test" given "disease"

$$= \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}}$$

- **Specificity**: probability of negative "test" given "healthy"

$$= \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}}$$

- **Recall** = true positive rate = sensitivity
- **Precision**: probability of "disease" given positive "test"

$$= \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}$$

Contigency table

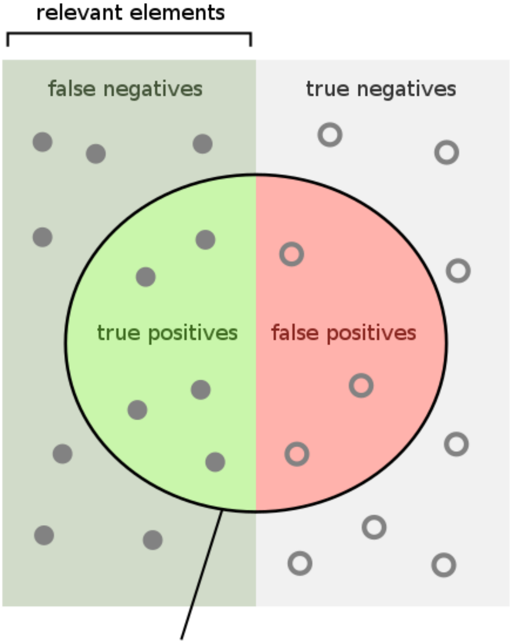
	bedroom count						
	0	1	2	3	4	5	6
low quality	28.57	44.44	52.48	51.19	48.47	57.45	57.14
high quality	71.43	55.56	47.52	48.81	51.53	42.55	42.86
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Confusion matrix

Self reported	Fact	High	Low	All
High		290	228	518
Low		202	280	482
All		492	508	1000

Contrast different goodness of fit measures

Sensitivity and specificity



Sensitivity and specificity

Bad ratios

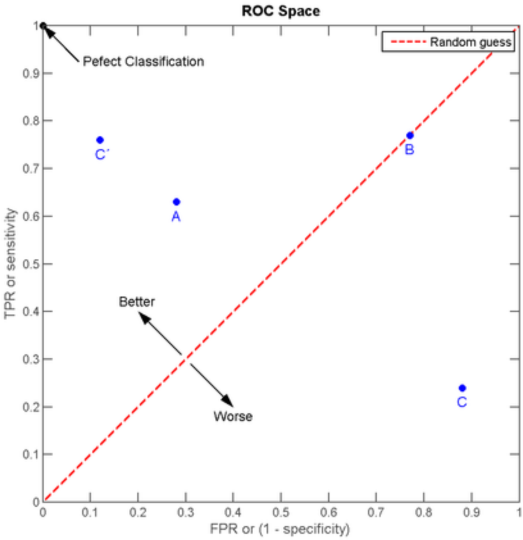
- False positive rate, Type-I error
- False negative rate, Type-II error

The ROC curve

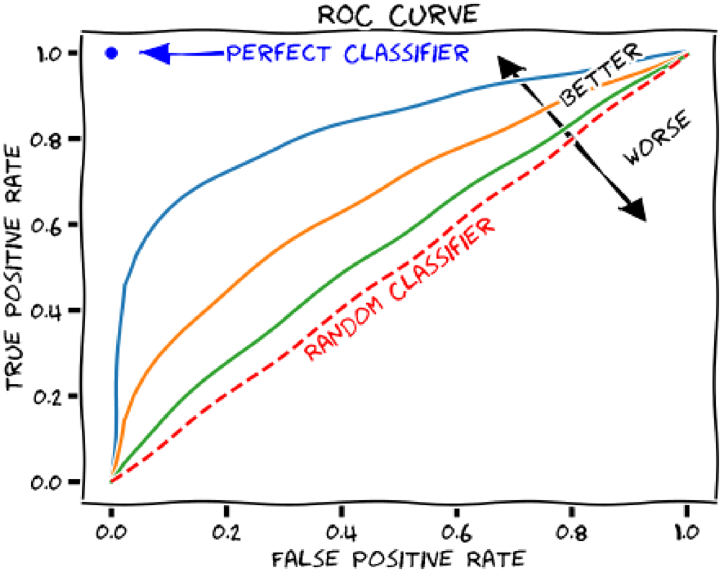
The trade-off

- You want to watch both types of errors. Otherwise it's easy to create a perfect prediction. How?
- Often we are trading off senstivity with specificity.
- Plot both on a graph (note the inverse scale) = **ROC curve** ("Receiver operating characteristic")

Different models



Different Parametrization of the Same Model



The area under the curve

A model often predicts an entire curve. An overall measure of performance is the **area under the curve** (AUC).

Properties

- Bounded between 0 and 1, higher means better fit.
- Symmetric in two types of error.
- Random chance (useless model): $AUC = 0.5$.

Contrast different goodness of fit measures

- Understand correlation, RMSE, R^2, AUC and confusion matrix.
- Relate type-I and type-II errors.

Discuss when ML improves decision making

Problem	Diagnostic	Improvements
Noisy prediction	Goodness of fit	Better (more) data, better model
Overfitting	Cross validation	Simpler model
Concept drift	Bad performance	Retrain model?
Covariate shift	Balance tests	Retrain model?
Wrong target metric	Insufficient lift	Select better metric
Non-actionable model	Now what?	Good questions first
Expensive deployment	$$$	Simpler data, model

Jargon busting

Regression

RMSE, correlation, R^2

Classification

confusion table, accuracy, false positive, false negative, sensitivity $=$ recall, specificity, precision, type-I and II error, ROC, AUC