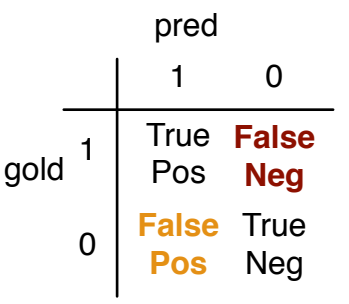
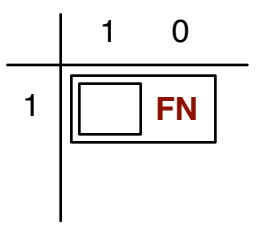
Confusion matrix metrics

These are conditional probabilities from counts on a binary confusion matrix.
Diagrams show the conditioning population and the quantity for the condprob numerator.
Odds notation X:Y≡X/(X+Y); and 1-(X:Y)=Y:X

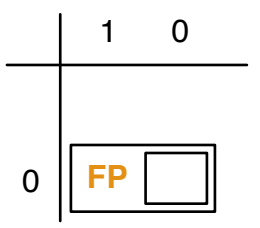
Brendan O'Connor
orig 2013-06-17, math fix 2015-03-12
http://brenocon.com/confusion_matrix_diagrams.pdf (.graffle)



Accuracy = p(correct) = (TP+TN) : (**FN**+**FP**)
But often you want to know about **FN**'s versus **FP**'s, thus all these metrics.

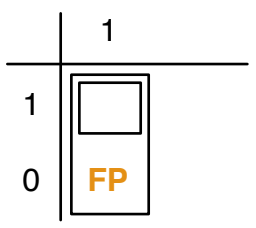


p(correct | gold=1)
Sensitivity a.k.a. **Recall**
a.k.a. TruePosRate
= TP : **FN**
= 1 - FalseNegRate
= 1 - (**FN**:TP)

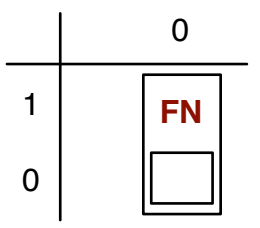


p(correct | gold=0)
Specificity
a.k.a. TrueNegRate
= TN : **FP**
= 1 - FalsePosRate
= 1 - (**FP**:TN)

Truth-conditional metrics
If you take the same classifier and apply it to a new dataset with different ground truth population proportions, these metrics will stay the same. Makes sense for, say, medical tests.



p(correct | pred=1)
Precision
a.k.a. Pos Predictive Value
= TP : **FP**
= 1 - FalseDiscoveryRate
= 1 - (**FP**:TP)



p(correct | pred=0)
Neg Predictive Value
= TN : **FN**

Prediction-conditional metrics
These are properties jointly of the classifier and dataset. For example,
p(person has disease | test says "positive")
= Precision = TP/(TP+FP)

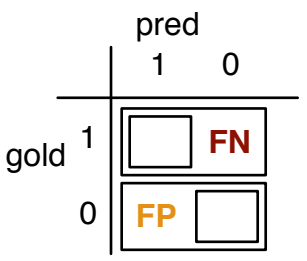
$$= \frac{\text{Sens}*\text{PriorPos}}{\text{Sens}*\text{PriorPos} + (1-\text{Spec})*\text{PriorNeg}}$$

Property of classifier Property of dataset

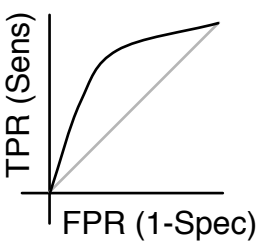
Confidence Tradeoffs

Classifier says YES if confidence in '1' is greater than a threshold *t*. Varying the threshold trades off between **FP**s versus **FN**s. For any one metric, you can make it arbitrarily good by being either very liberal or very conservative. Therefore you are often interested in a pair of metrics: one that cares about **FP**s, and one that cares about **FN**s. Here are two pairings that are popular. By changing the confidence threshold, you get a curve of them.

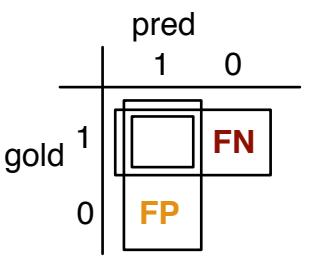
Sens-Spec Tradeoff: **ROC curve**
(sens, spec) = (TP:**FN**, TN:**FP**)



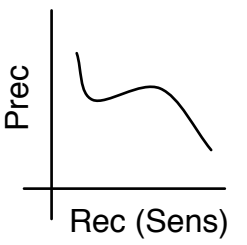
Ignores truth marginals (row proportions)



Sens-Prec Tradeoff: **Precision-Recall curve**
(prec, rec) = (TP:**FP**, TP:**FN**)



Ignores true negatives: makes sense for needle-in-haystack problems, where TNs dominate and are uninteresting.



There are several summary statistics to make a **FP**-**FN** metric pairing invariant to the choice of threshold.

AUC *area under the ROC curve*. (a.k.a. ROCAUC, AUROC)
Semantics: prob the classifier correctly orders a randomly chosen pos and neg example pair. (Similar to Wilcoxon-Mann-Whitney or Kendall's Tau). Neat analytic tricks with convex hulls.

PRAUC: *area under the PR curve*.
No semantics that I know of.
Similar to **mean average precision**. See also **precision@K**.

PR Breakeven: find threshold where prec=rec, and report that value.

Alternatively, directly use predicted probabilities without thresholding or confusion matrix:
Cross-entropy (log-likelihood) or **Brier score loss** (squared error).

There are also summary stats that require a specific threshold -- or you can use if you don't have confidences -- but have been designed to attempt to care about both FP and FN costs.

Balanced accuracy: arithmetic average of sens and spec.
This is accuracy if you chose among pos and neg classes with equal prob each -- i.e. a balanced population.
[In contrast, AUC is a balanced *rank-ordering* accuracy.]
Corresponds to linear isoclines in ROC space.

Balanced accuracy = macroaveraged recall

Accuracy = microavg rec = microavg prec = microavg f1

F-score: harmonic average of prec and rec.
Usually **F1** is used: the unweighted harmonic mean.
F1 = 2*P*R / (P+R)
No semantics that I know of.
Corresponds to curvy isoclines in PR space.

Multiclass setting: each class has it own prec/rec/f1.
Macroaveraged F1: mean of each class' F1.
Cares about rare classes as much as common classes.
Microaveraged F1: take total TP,FP,FN to calculate prec/rec/f1.
Cares more about highly prevalent classes.

Classical Frequentist Hypothesis Testing

Task is to detect null hypothesis violations.
Define "reject null" as a positive prediction.
- Type I error = **false positive**. **Size (α)** = 1-Specificity
- Type II error = **false negative**. **Power (β)** = Sensitivity

and **p-value** = 1-Specificity, kind of.
"Significance" refers to either p-value or size (i.e. **FP**R)

The point is to give non-Bayesian probabilistic semantics to a single test, so these relationships are a little more subtle. For example, size is a worst-case bound on specificity, and a p-value is for one example: prob the null could generate a test statistic at least as extreme (= size of the strictest test that would reject the null on that example).
When applied to a single test, PPV and NPV are not frequentist-friendly concepts since they depend on priors. But in multiple testing, they can be frequentist again; e.g. Benjamini-Hochberg bounds expected PPV by assuming PriorNeg=0.999999.

References

Diagrams based on William Press's slides: http://www.nr.com/CS395T/lectures2008/17-ROCPrecisionRecall.pdf
Listings of many metrics on: http://en.wikipedia.org/wiki/Receiver_operating_characteristic
Fawcett (2006), "An Introduction to ROC analysis": http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf
Manning et al (2008) "Intro to IR", e.g. http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html
Hypothesis testing: Casella and Berger, "Statistical Inference" (ch 8); also Wasserman, "All of Statistics" (ch 10)