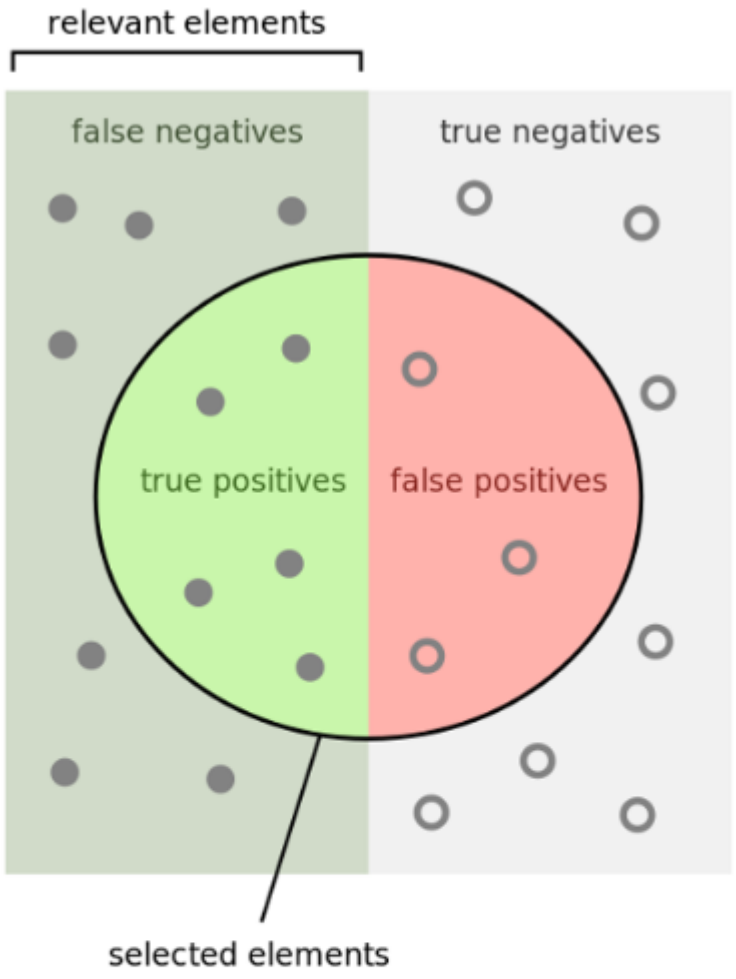
Precision and recall

https://en.wikipedia.org/wiki/Precision_and_recall (https://en.wikipedia.org/wiki/Precision_and_recall)
https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics (https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics)

People often talk about the same error metrics with very different names, as in the information retrieval or classification context. To be crystal clear about these confused names, it is essential to understand and to be very familiar with the following figure (Wikipedia). The vertical middle line indicates an balanced case where y=0 and y=1 have similar number of elements. For a skewed class, **the line may move to the far left or right.**



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Be very familiar with the components of the above figure. Almost all the concepts can be understood with this figure.

- Understand the 'relevant elements' as the y=1 class such as cancer positive, spam email etc.
- Understand the 'selected elements' as all the predicted y=1 elements, or predicted positive elements. Sometimes they are also called retrieved elements.
- Left half describes all y=1 relevant elements, and the right half describes all y= 0 elements.
- The elements in the left half but within left semi-circle are predicted y=1 elements, i.e true positive (TP).
- The elements in the left half but outside the left semi-circle are false negative (FN).
- The elements in the right half but within right semi-circle are predicted y=1 elements, i.e false positive (FP).
- The elements in the right half but outside the right semi-circle are true negative (TN).

After clarifying these concepts, we define the following error metrics.

- recall = TP/(TP+FN), also referred to as 'true positive rate' or 'sensitivity'.
  - TP+FN are all the elements of y=1 class, or the relevant elements. So 'recall' refers to how many y=1 elements we RECALLED (found, retrieved) from all the y=1 elements.
  - From the figure, we know that the higher the recall (up to 1), the better. However, even recall = 1 does not necessarily indicate a best model. This is because if the model also predicted a lot of false positives, then it is still not good, even recall = 1. So we need look at other metrics such as precision.
- precision = TP/(TP+FP), also referred to as 'positive predictive value'.
  - How precise is my predicted y=1 class elements? That is, among all the predicted y=1 elements (TP+FP), what fraction is the true positives?
  - From the figure, we know the higher the precision (up to 1), the better. However, even precision = 1 does not necessarily indicate a best model. Precision = 1 suggests the whole circle is located to the left. But if the circle is relatively small as compared to the left half, then we have a lot of false negatives.
- accuracy = (TP+TN)/(TP+TN+FP+FN)
  - Accuracy does not perform well with imbalanced data sets. For example, if you have 95 negative and 5 positive samples, classifying all as negative gives 0.95 accuracy score. Balanced Accuracy overcomes this problem, as below.
- balanced accuracy = (TPN+TPR)/2, see details in Wikipedia.
  - obtained by normalizing true positive and true negative predictions by the number of positive and negative samples, respectively, and divides their sum into two.
  - Regarding the previous example (95 negative and 5 positive samples), classifying all as negative gives 0.5 balanced accuracy score out of the maximum bACC one, which is equivalent to the expected value of a random guess of a balanced data. Balanced Accuracy is suggested to use to measure how accurate is the overall performance of a model is, considering both positive and

negative classes without worrying about the imbalance of a data set. Since most of the real data sets are imbalanced, **Balanced Accuracy metric is suggested instead of Accuracy metric.**

- Many other indicators. See the Wikipedia link.

Trading off precision and recall

Case 1:
Predict 1 if hθ(x) >= 0.5. Predict 0 if hθ(x) < 0.5. This classifier may give some value for precision and some value for recall. Understand this from the normal recall-precision figure.

Case 2: Predict 1 if hθ(x) >= 0.8. Predict 0 if hθ(x) < 0.2. Now we can be more confident a 1 is a true positive. But classifier has lower recall - predict y = 1 for a smaller number of patients. Risk of false negatives. Understand this from the recall-precision figure but with a **shrunk circle** moving to the left (meaning less false positive).

Case 3:
Predict 1 if hθ(x) >= 0.3. Predict 0 if hθ(x) < 0.7. Set a lower threshold. In the cancer example, we have have a higher recall, but lower precision. Risk of false positives, because we're less discriminating in deciding what means the person has cancer. Understand this from the recall-precision figure but with an **expanded circle** in the middle.

This threshold defines the trade-off as below. **From the figure above make sure be familiar with the three cases above. Although there is usually a trade off between recall and precision, this is not always the case. For a perfect classifier, both recall and precision can be 1. This corresponds to the case where the circle in the middle will reshape and move to the left to be exactly congruent with the left half of the figure.**

Single or multiple error metrics?

A single error metric such as accuracy is not enough to handle all the cases such as skewed class case. Therefore we introduce recall and precision. These two indicators can sometimes useful in many cases. For example, in different cases such as cancer detection, etc., we may either favor recall or precision. However, sometimes it is hard to determine which algorithm is better if we compare their (recall, precision) pairs. To easily compare algorithms, it is preferable to design some single real number metric. As already introduced earlier, the balanced accuracy designed for skewed class is one of them. Next we introduce a few more such indicators.

- Balanced accuracy (see definition earlier or in Wikipedia).
- F1 score or fscore = 2PR/(P + R)
  - If P = 0 or R = 0 the fscore = 0. If P = 1 and R = 1 then Fscore = 1. The remaining values lie between 0 and 1.

- Threshold offers a way to control trade-off between precision and recall. Fscore gives a single real number evaluation metric. If you're trying to automatically set the threshold, one way is to try a range of threshold values and evaluate them on your cross validation set. Then pick the threshold which gives the best fscore.

Type I and Type II errors.

https://en.wikipedia.org/wiki/Type_I_and_type_II_errors (https://en.wikipedia.org/wiki/Type_I_and_type_II_errors)
Type I and Type II errors are related to the recall-precision figure shown earlier.

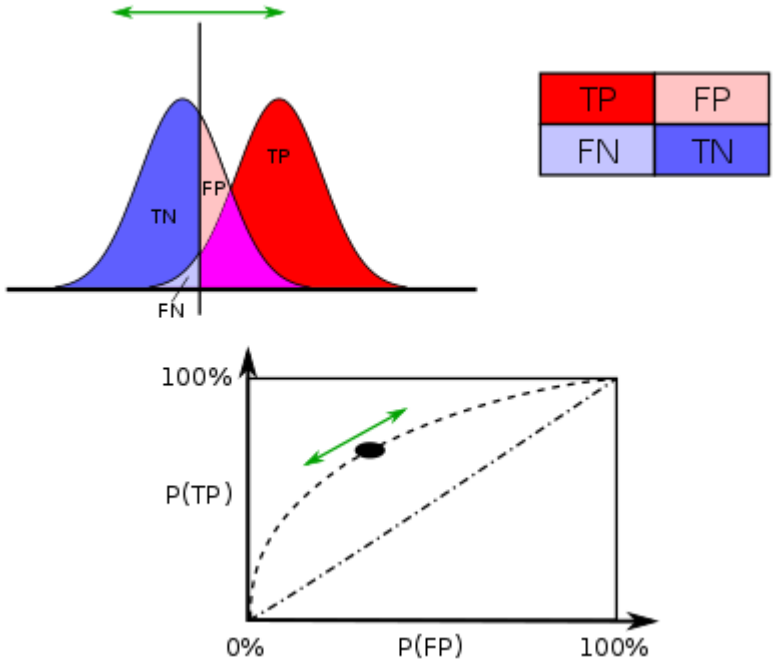
- A type I error is a false positive (half-circle to the right of figure), and a type II error is a false negative (outside the half circle to the left). Comment: if we plot the recall-precision figure with y = 0 in the left, the type I elements are in the left. If we plot as the figure shown earlier, then type I element is in the right. This later case is often shown in a matrix-like form.
- Type I error is the incorrect rejection of a **true null hypothesis** (the elements should be in the right of figure y=0, or negative, but we reject). Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't. Use False Positive to understand the following examples.
  - A test that shows a patient to have a disease when in fact the patient does not have the disease.
  - A fire alarm going on indicating a fire when in fact there is no fire.
  - An experiment indicating that a medical treatment should cure a disease when in fact it does not.
- A type II error is incorrect rejection of a **false null hypothesis** (the elements should be in the left of figure y=1, or positive, but we reject). Use False Negative to understand the following examples.
  - A blood test failing to detect the disease it was designed to detect, in a patient who really has the disease;
  - A fire breaking out and the fire alarm does not ring;
  - A clinical trial of a medical treatment failing to show that the treatment works when really it does.
- **A way to memorize type I and type II**. Mapping type I with type 0 (y=0) and type II with type 1 (y=0). So type I is reject y= 0 case (Predict disease when there is none), and type II is rejecting y = 1 (Predict no disease when there is disease).

Receiver operating characteristic (ROC)

https://en.wikipedia.org/wiki/Receiver_operating_characteristic (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)
After a deep understanding of the concepts such as recall and precision and the intuitive figure, it is straightforward to understand their derivatives such as ROC curve. After understanding ROC curve, then we can also understand why people can use Area Under Curve (AUC) to select variables.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. **TPR is just the recall defined earlier, also called sensitivity.**

For example, imagine that the blood protein levels in diseased people and healthy people are normally distributed with means of 2 g/dL and 1 g/dL respectively. A medical test might measure the level of a certain protein in a blood sample and classify any number above a certain threshold as indicating disease. The experimenter can adjust the threshold (black vertical line in the figure below), which will in turn change the false positive rate. Increasing the threshold would result in fewer false positives (and more false negatives), corresponding to a leftward movement on the curve. The actual shape of the curve is determined by how much overlap the two distributions have.



When shifting the threshold, the circle (not necessarily a real circle) in the recall-precision figure will expand or shrink, and the center of the circle will also shift either to the left or right. This gives the varying of TP and FP. **Mapping the sliding of cutoff line of the recall-precision figure with the ROC curve, or with the pink overlapping picture above**.

Coefficient of determination

https://en.wikipedia.org/wiki/Coefficient_of_determination (https://en.wikipedia.org/wiki/Coefficient_of_determination)
In statistics, the coefficient of determination, denoted R^2 and pronounced "R squared", is **the proportion of the variance in the dependent variable that is predictable from the independent variable(s)**.

Definitions

A data set has n values y_i where $i = 1, 2, \ldots n$, each associated with a predicted value f_i. The residuals is defined as $e_i = y_i - f_i$. Also define $\bar{y} = \frac{1}{n}\sum_i^n y_i$. Then the variability of the data can be measured using THREE sums of squares formulas.

- The total sum of squares
$$SS_{tot} = \sum_i (y_i - \bar{y})^2,$$
which is proportional to the variance of the data.
- The regression sum of squares, also known as explained sum of squares
$$SS_{reg} = \sum_i (f_i - \bar{y})^2$$
- The sum of squares of residuals, also known as residual sum of squares
$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$$

Check the figure showing the meaning of R^2 in the Wikipedia page.

Note as sample mean, χ^2, etc., here R^2 is just another statistic defined.

Relation to unexplained variance

$$R^2 = 1 - FVU$$

where FVU is the fraction of the variance unexplained.

As explained variance

In some cases such as OLS model, we have $SS_{res} + SS_{reg} = SS_{tot}$. The we have

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{SS_{reg}/n}{SS_{tot}/n}$$

where n is the number of observations. In this form R^2 is expressed as the ratio of the explained variance to the total variance.

As squared correlation coefficient

In linear least squares multiple regression with an estimated intercept term, R^2 equals the square of the Pearson correlation coefficient between the observed y and modeled (predicted) f data values of the dependent variable.

In a linear least squares regression with an intercept term and a single explanator, this is also equal to the squared Pearson correlation coefficient of the dependent variable y and explanatory variable x.

Range of R^2

Values of R^2 outside the range 0 to 1 can occur when the model fits the data worse than a horizontal hyperplane. This would occur when the wrong model was chosen, or nonsensical constraints were applied by mistake.
However, generally R^2 will be within the range 0 to 1, and 1 indicates regression predictions perfectly fit the data.

Inflation of R^2

In least squares regression, R^2 is weakly increasing with increases in the number of regressors in the model. Because increases in the number of regressors increase the value of R^2, R^2 alone cannot be used as a meaningful comparison of models with very different numbers of independent variables. For a meaningful comparison between two models, an F-test can be performed on the residual sum of squares, similar to the F-tests in Granger causality, though this is not always appropriate.

Adjusted R^2

The use of an adjusted R^2 (one common notation is $\bar{R}^2$ and another is R^2_{adj} is an attempt to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model. The adjusted R^2 can be negative, and its value will always be less than or equal to that of R^2. Unlike R^2, the adjusted R^2 increases only when the increase in R^2 (due to the inclusion of a new explanatory variable) is more than one would expect to see by chance. The adjusted R^2 is defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where where p is the total number of explanatory variables in the model (not including the constant term), and n is the sample size.

Variance inflation factor (VIF) and tolerance

https://en.wikipedia.org/wiki/Variance_inflation_factor (https://en.wikipedia.org/wiki/Variance_inflation_factor)
The variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. **It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.** It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

See the link above about the steps of calculating VIF. **Tolerance is just defined as 1/VIF.**

Comments: It seems VIF or tolerance can be replaced by correlation in the analysis multicollinearity.