Evaluating a Model

We'll cover the following

Precision, Recall, and Confusion Matrix

We have learned about various ML models, but how do we evaluate them? For regression, we can use the difference between the actual and the predicted values — Root Mean Square Error, RMSE, or ordinary least square method, to be more precise — but what about classification models?

One might think that accuracy is a good enough measure to evaluate the goodness of a model. Accuracy is a very important evaluation measure, but it might not be the best metric all the time. Let's understand this with an example.

The Accuracy Trap

Say we are building a model that predicts if patients have a chronic illness. We know that only 0.5% of the patients have the disease, or are "Positive" cases. Now, a dummy model could always give "Negative" as a default result and still have a high accuracy (99.5%!) because our dataset is skewed. Out of all the patients only 0.5% have the disease, so by giving "Negative" as a default answer for 100% of the cases, the model is still able to get the predictions right in 99.5% of the cases – we have a model with a very high accuracy! But is this of any good? Absolutely not! And this is where some other performance measures come into play.

Precision, Recall, and Confusion Matrix

Before we talk about these measures, let's understand a few terms:

1. **TP / True Positive**: the case was positive, and it was predicted as positive
2. **TN / True Negative**: the case was negative, and it was predicted as negative
3. **FN / False Negative**: the case was positive, but it was predicted as negative
4. **FP / False Positive**: the case was negative, but it was predicted as positive

Since pictures help us to remember things better:

Image Credits: http://www.info.univ-angers.fr

Now that we know the meaning of false positives, false negatives, true positives, and true negatives, we can learn about the famous **Confusion Matrix**.

A confusion matrix has two rows and two columns that report the number of false positives, false negatives, true positives, and true negatives. Basically, it is a summary table showing how good our model is at predicting examples of various classes.

For example, if we have a classification model that has been trained to distinguish between cats and dogs, a confusion matrix will summarize the

results of testing the algorithm on new data. Assuming a sample of 13 animals — 8 cats and 5 dogs — our confusion matrix would look like this:

		Actual class	
		Cat	**Dog**
Predicted class	**Cat**	**5**	2
	Dog	3	**3**

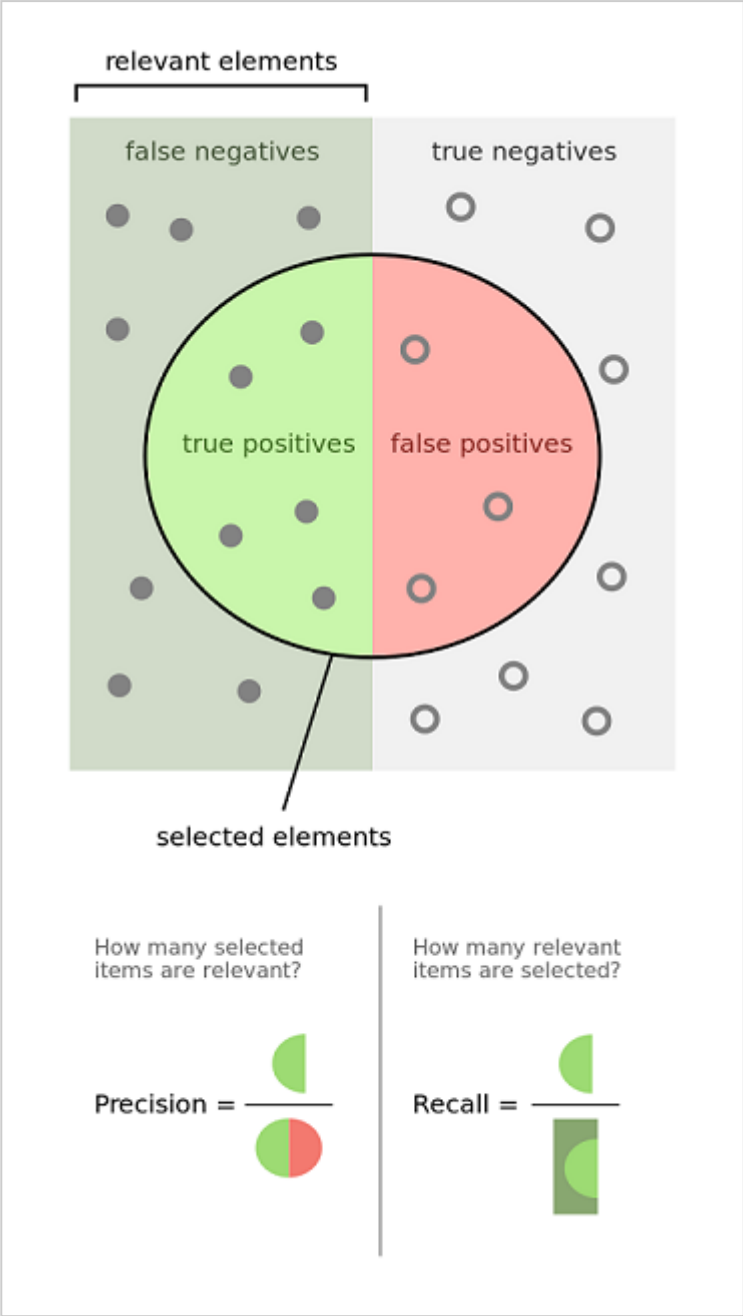
Based on this, we can obtain two important measures:

- **Precision**: The ratio of correct positive predictions to the total ***predicted positives***, the ***positive predictive value***

  $$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall**: Ratio of correct positive predictions to the ***total actual positives*** examples in the dataset, the ***sensitivity***

  $$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



Precision and recall. Image Credits: Wikipedia

Putting this all together, which would be the correct measure to answer the following questions?

1. ***What percentage of our predictions were correct?***

- Accuracy

2. ***What percentage of the positive cases did we identify?***

- Recall

3. ***What percentage of positive predictions were correct?***

- Precision

In our case of predicting if a person has a chronic illness, it would be better to have a high Recall because we do not want to leave any untreated any patients who have the disease. It's better to have false alarms rather than missing positive cases, so we might be okay with the **low precision but high recall trade-off**.

Note: In case our dataset is not skewed, but rather a balanced representation of the two classes, then it is totally okay to use **Accuracy** as an evaluation measure:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Worked Example

Before continuing on, let's look at a completed example:

Suppose the fecal occult blood (FOB) screen test is used in 2030 people to look for bowel cancer:

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Image Credits: Wikipedia

AUC-ROC Curve

AUC (Area Under the Curve) - ROC (Receiver Operating Characteristics) curve is a performance measurement for a classification model at various classification threshold settings. Basically, it is a probability curve that tells us how well the model is capable of distinguishing between classes. The higher

the AUC value of our probability curve, the better the model is at predicting 0s as 0s and 1s as 1s.

What do we mean by various threshold settings?
Say we set the threshold to 0.9. This means that if for any given sample our trained model predicts a value higher than 0.9, our output class will be predicted as positive class; otherwise, it will be placed in the negative class.

The ROC curve is plotted with True Positive Rate (Recall/Sensitivity) against the False Positive Rate (FPR, *1 - Specificity*) where TPR is on y-axis and FPR is on the x-axis, where:

- **Sensitivity, Recall, Hit Rate, or True Positive Rate (TPR)**

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Fall-out, (1 - Specificity) or False Positive Rate (FPR)**

$$1 - \text{Specificity} = \text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

A great model has AUC near the 1 indicating it has an excellent measure of separability. On the other hand, a poor model has AUC near to the 0 meaning it is predicting 0s as 1s and 1s as 0s. And When AUC is 0.5, it means the model has no class separation capacity whatsoever and it's essentially making random predictions.

Let's understand this better via an example analysis taken from a medical research journal:

ROC Curve Analysis: Example Case Study

We've identified a potential biomarker, *Protein "A"*, of Alzheimer's disease that is elevated in Alzheimer's patients compared to healthy patients (Figure 1 in the image below). We now need to identify a good threshold value of this protein in order to have a model that can be used to identify Alzheimer's patients with a good performance.

- **If the threshold is too low**: a lot of healthy patients will be wrongly diagnosed

- **If the threshold is too high**: a lot of healthy patients will be wrongly diagnosed

A ROC curve can help us in ***identifying the sweet spot***, a balance between TPR and FPR.

A ROC curve is generated across all the threshold settings and the AUC (area under the curve) value is determined (Figure 3 in the image below).

- Higher AUC values indicate a better biomarker.
- For our final mode, we choose a point along the ROC curve (value for the threshold) so that we have an acceptable or optimal **trade-off between sensitivity and specificity**.

Let's say the black dashed line is the ROC curve for our data in this example. We could choose $X = 0.1$ and $Y = 0.8$, so that our model based on the given biomarker, *Protein A*, would have a specificity of 90% and a sensitivity of 80% in identifying Alzheimer's patients.

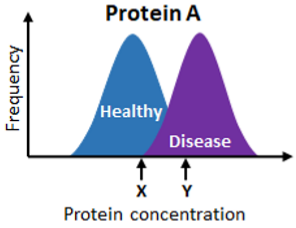


Figure 1. Overlapping histogram plots for concentrations of Protein A in different populations. A cut-off of concentration "X" will have high sensitivity, but low specificity. A cut-off of concentration "Y" will have low sensitivity, but high specificity.

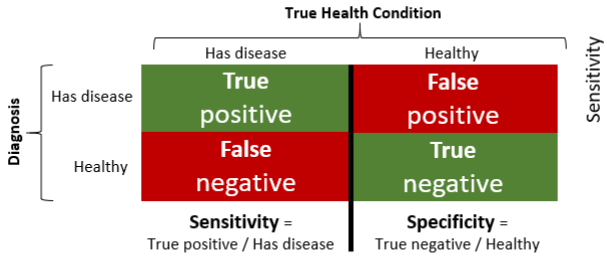


Figure 2. Calculation of sensitivity and specificity

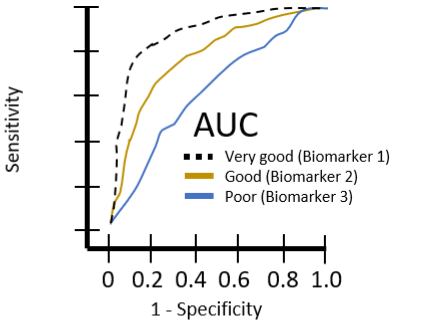


Figure 3. Comparison of ROC curves across three potential biomarkers. The higher the AUC value, the higher predictive value of the biomarker. Biomarker 3 has very poor predictive power (AUC ~0.5) as it cannot differentiate between healthy and diseased patients at all.

ROC-Curve-Analysis. Image Credits: https://raybiotech.com/learning-center/roc-curve-analysis/