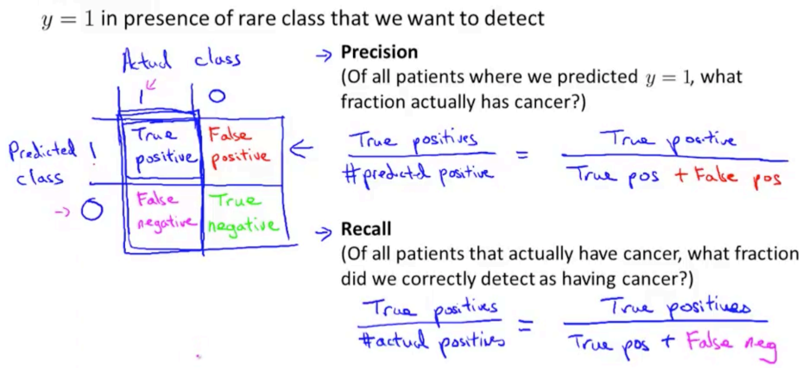
Skewed Dataset - Large Dataset

Handling Skewed Dataset

Sometimes we encounter a skewed dataset. For example, our dataset shows only 98% of patients have cancer (y=1) and the rest do not have cancer (y=0). If we apply our model, we may get 99% accuracy, but this is a wrong accuracy. Even if your model by default assume no cancer all the time, the model will have 98% accuracy! In this case, we may use precision recall as a new metric for measuring the performance of the model.

Precision Recall

The figure below shows a matrix consist of false negative, true negative, rue positive, and false positive.



$y = 1$ in presence of rare class that we want to detect

		Actual class	
		1	0
Predicted class	1	True positive	False positive
	0	False negative	True negative

→ **Precision**

(Of all patients where we predicted $y = 1$, what fraction actually has cancer?)

$$\frac{\text{True positives}}{\text{\#predicted positive}} = \frac{\text{True positive}}{\text{True pos} + \text{False pos}}$$

→ **Recall**

(Of all patients that actually have cancer, what fraction did we correctly detect as having cancer?)

$$\frac{\text{True positives}}{\text{\#actual positives}} = \frac{\text{True positives}}{\text{True pos} + \text{False neg}}$$

Precision indicates how often algorithm cause a false alarm and it's defined by how many patients we predicted have cancer vs fraction of how many of them actually have cancer. High precision is good (i.e. closer to 1):

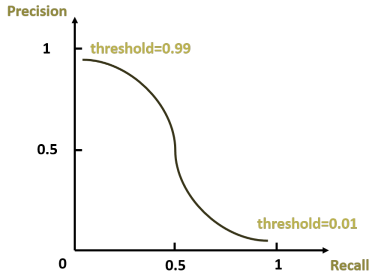
- We want a big number, because we want false positive to be as close to 0 as possible

Recall indicates how sensitive our algorithm is. It is defined by of all patients who have cancer, what fraction we detect correctly. High recall is good (i.e. closer to 1)

- We want a big number, because we want false negative to be as close to 0 as possible

It's pretty hard to look at 2 numbers for evaluation, therefore we define f1-score for ease in model evaluation.

Precision Recall Trade off- F1-Score



For many applications, we need to control the tradeoff between precision and recall. For example, in cancer example we want to avoid false negative so we predict 1 if h(x) > 0.3. In this case we have high recall and lower precision (avoid missing too many cases of cancers). In some other cases, we may care about lower recall. The graph in the right shows the precision recall tradeoff. Having a single value for evaluation metrics helps us decide easier. F1 score gives us single value for evaluation and is defined as:

$$\boxed{\text{F1 score} = 2 * \frac{\text{PR}}{\text{P} + \text{R}}}$$

Skewed Dataset - Large Dataset

F1 score is like taking average between P and R and gives higher rate to the lower value. Let’s go over few case scenario here:
If P=0 or R=0 then F1 score is equal to 0
If P=1 and R=1 then F1 score is equal to 1
Else F1 score is value between 0 and 1

In order to automatically set the threshold for hypothesis, 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 F1 score.