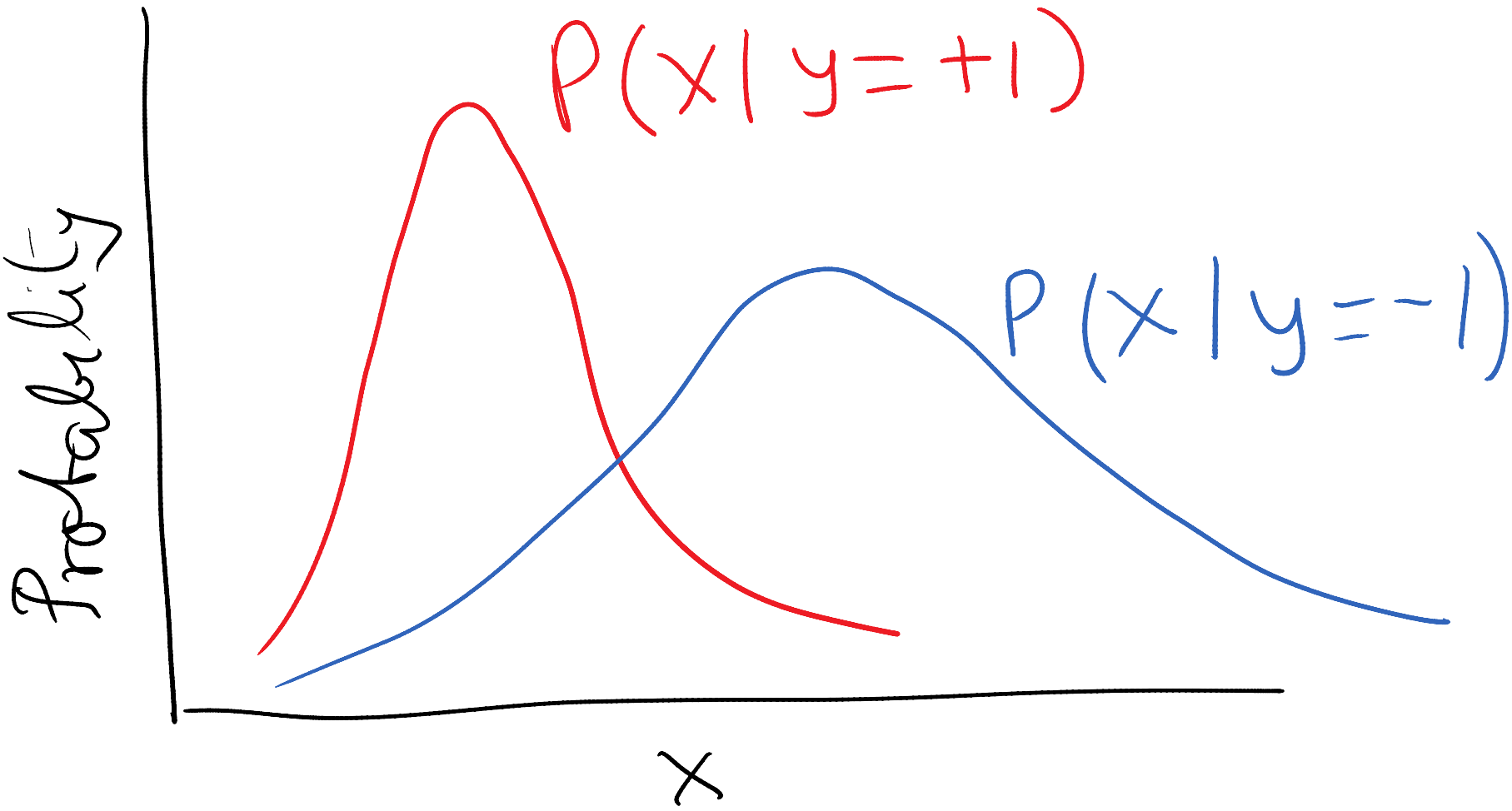
Decision Theory for Classification

In general, the feature vector doesn't uniquely specify the class. Therefore the best we can do is to guess, and we will make some mistakes. But what is the best we can do?

Given x, what should we guess for y



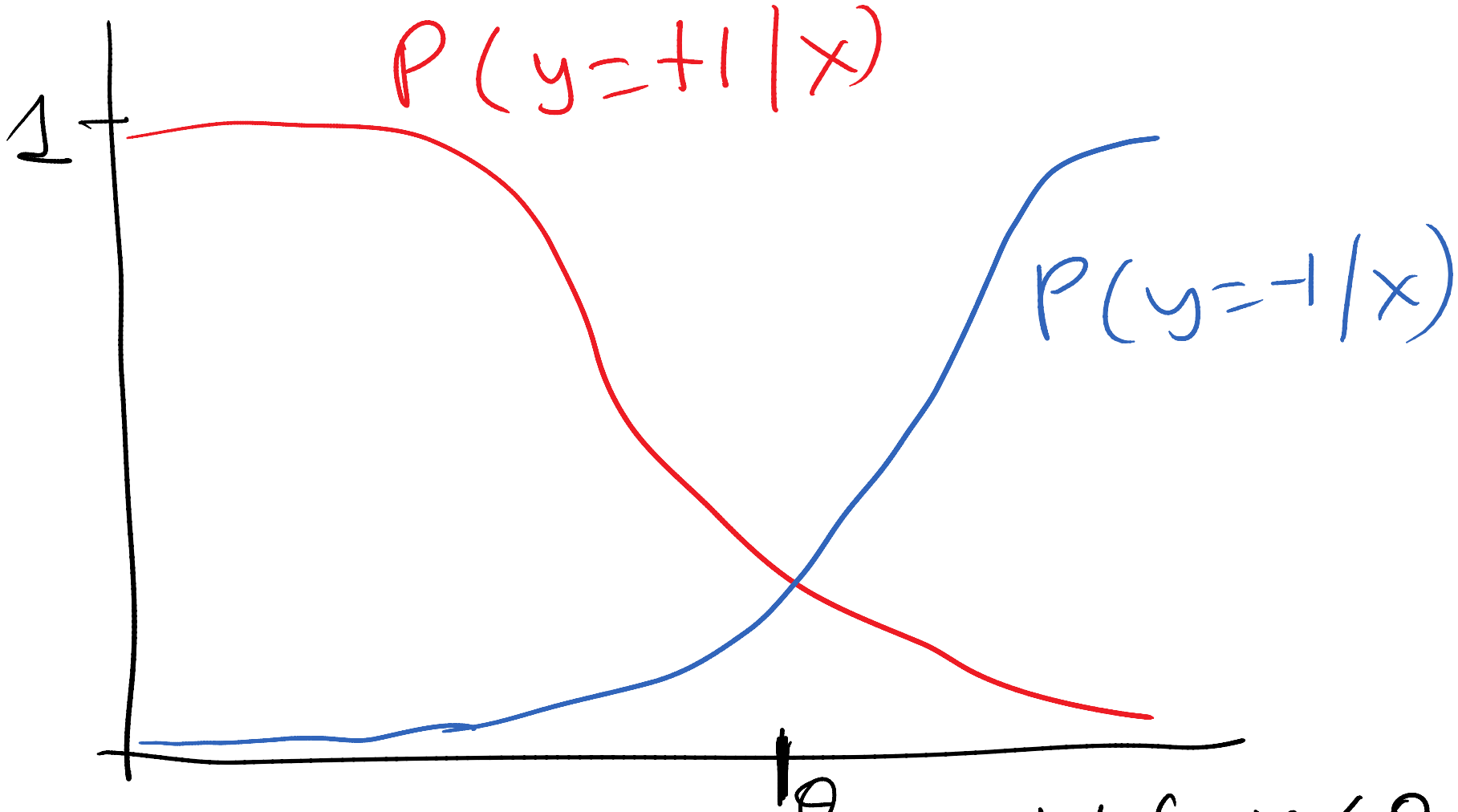
Use Bayes Rule

$$P(y=+1 \mid x) = \frac{P(x \mid y=+1)\,P(y=1)}{P(x)}$$

$$P(y=-1 \mid x) = \frac{P(x \mid y=-1)\,P(y=-1)}{P(x)}$$

Suppose $P(y=1) = \frac{2}{3}$, $P(y=-1) = \frac{1}{3}$

Posterior probabilities



So should we guess $y = +1$ for $x < \theta$
-1 for $x > \theta$

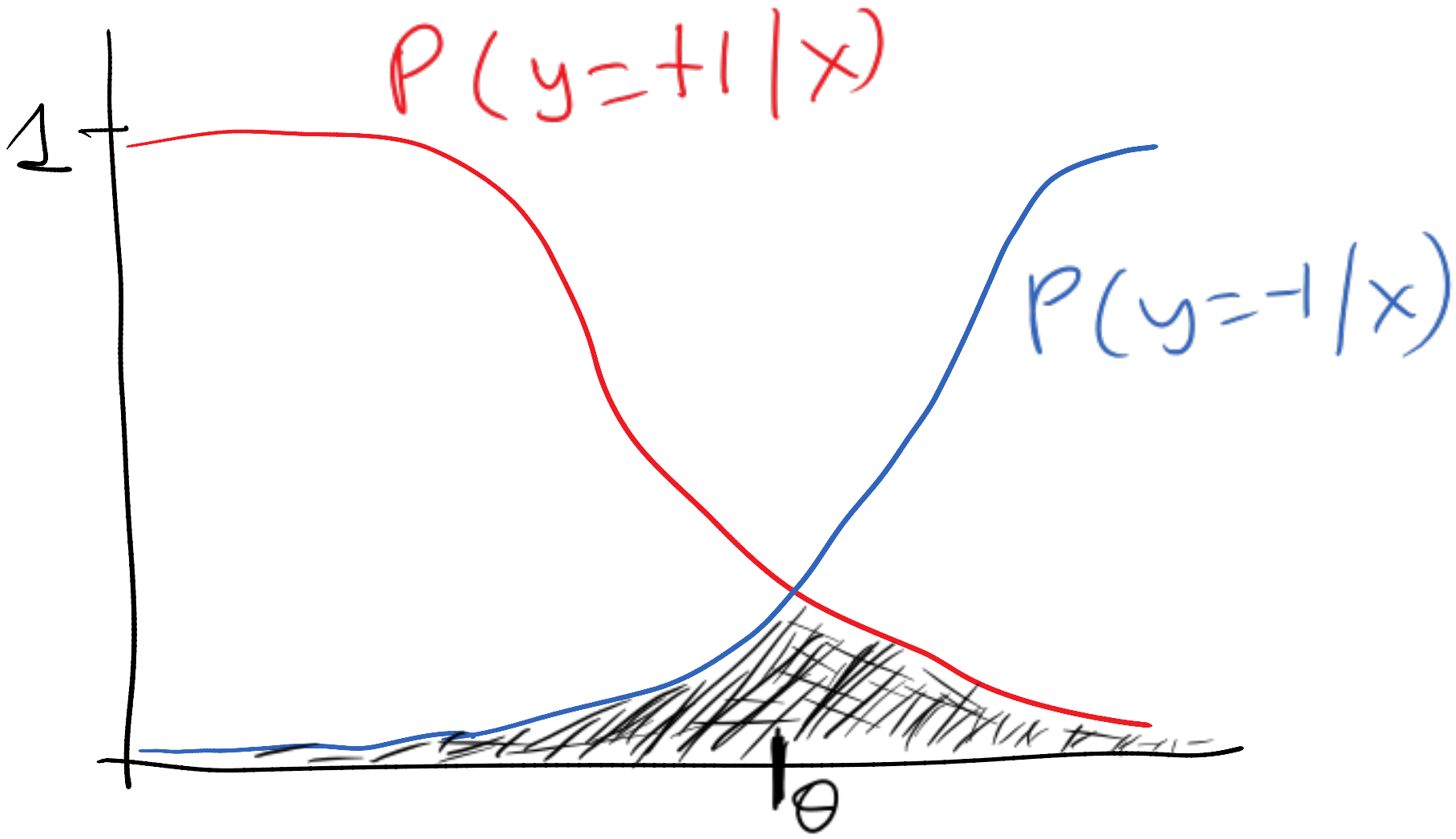
Depends on the loss function!

- Yes, if the goal is to minimize the probability of misclassification

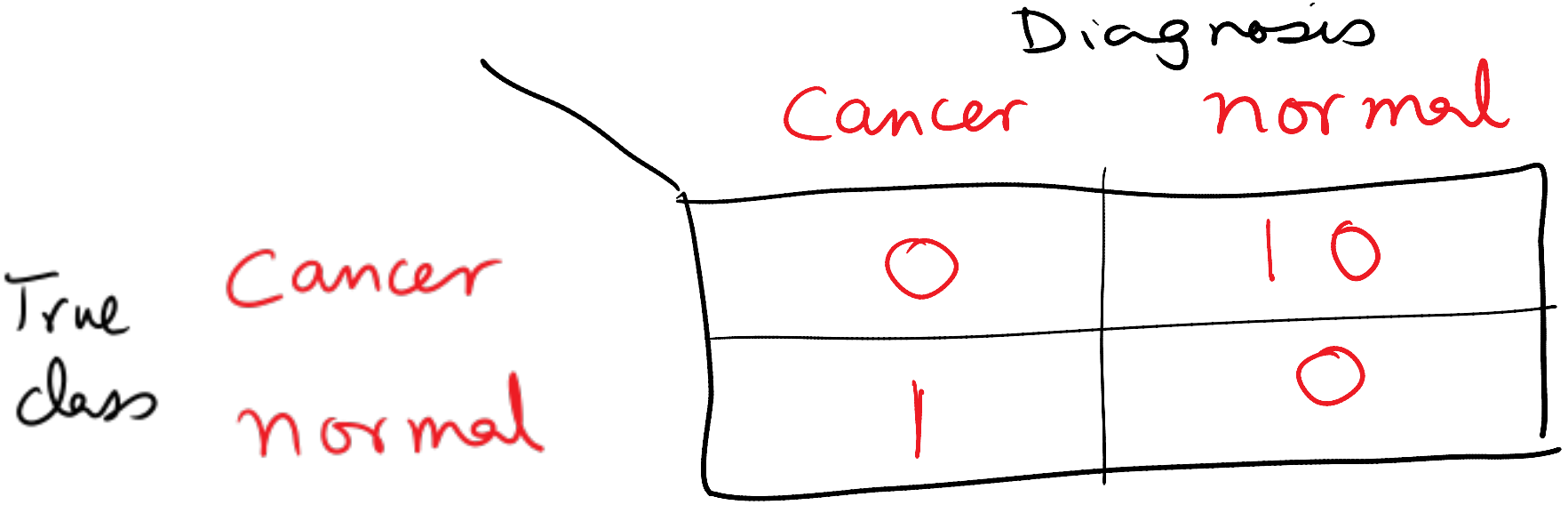
$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}|x)\,P(x)\,dx$$

To minimize $P(\text{error}|x)$ choose class with higher posterior probability

Bayes Risk $= \int_{-\infty}^{\infty} P(\text{error}|x)\, P(x)\, dx$



Sometimes we have an asymmetric loss function



To minimize expected loss, we will err on the side of diagnosing cancer for normals, rather than the other way around.

Modified Rule

L_{kj} is loss when true class is k, but we declare j

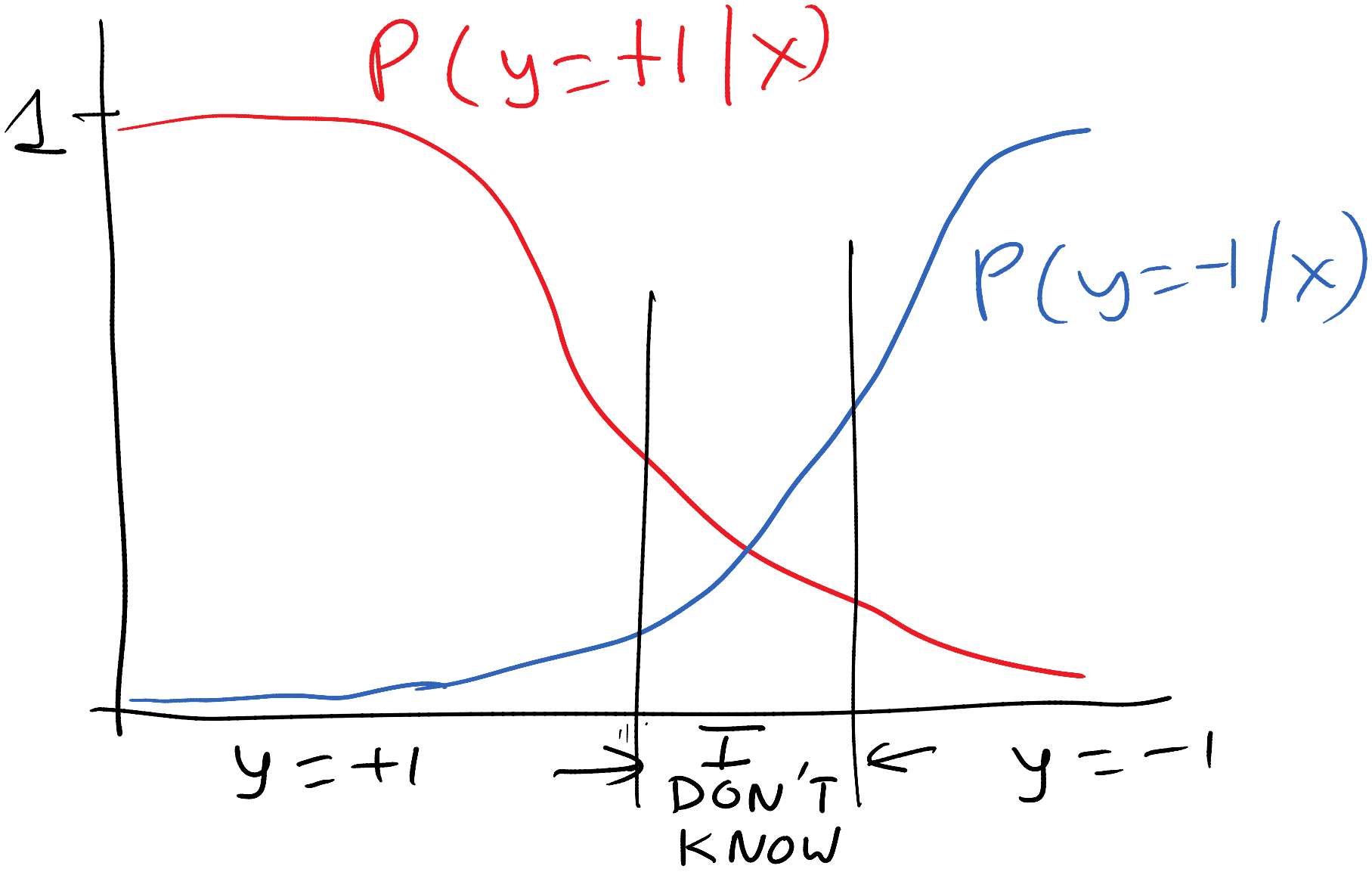
Declare class j for which

$$\sum_{K} L_{Kj}\, P(C_K|x)$$

is a minimum

By integrating over x, we can compute the expected Loss

Using the "Doubt" option



Good classifiers vs. Bad classifiers

- Good classifiers have an expected loss that is closer to the Bayes Risk of the classification problem, given a certain choice of features.

Three ways of building classifiers

- Generative
- Discriminative
- Find decision boundaries

. Model $P(\underset{\sim}{x} | C_k)$
. Model $P(C_k)$
obtain $P(C_k | \underset{\sim}{x})$ using Bayes Theorem

. Model $P(C_k | \underset{\sim}{x})$

. Model $f: \underset{\sim}{x} \rightarrow k$

Bayes Optimal Classifier

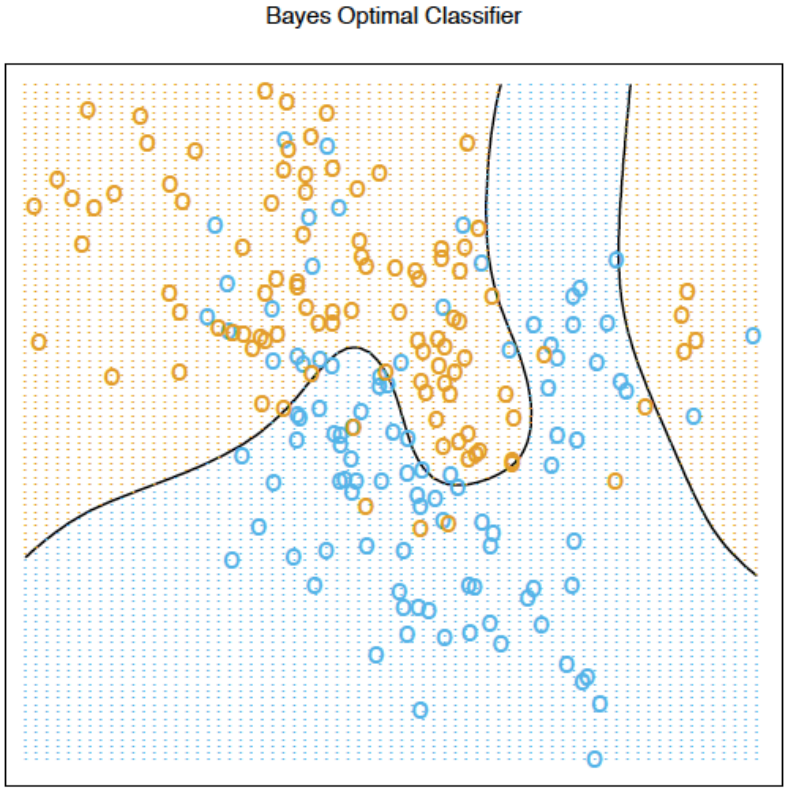


FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

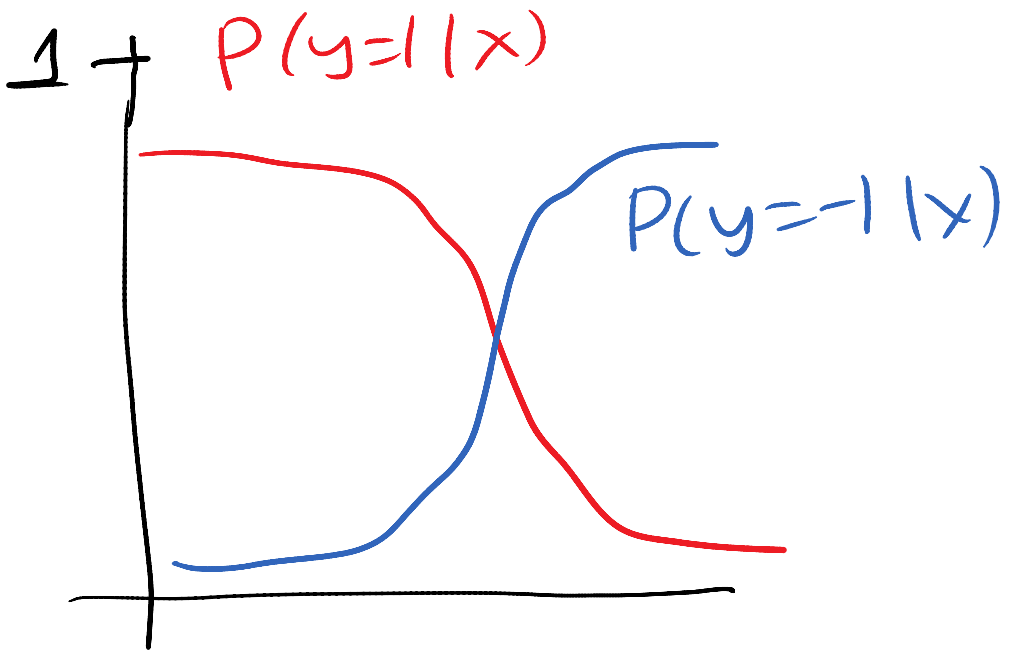
How the data was generated (ESLI, Chapter 2)

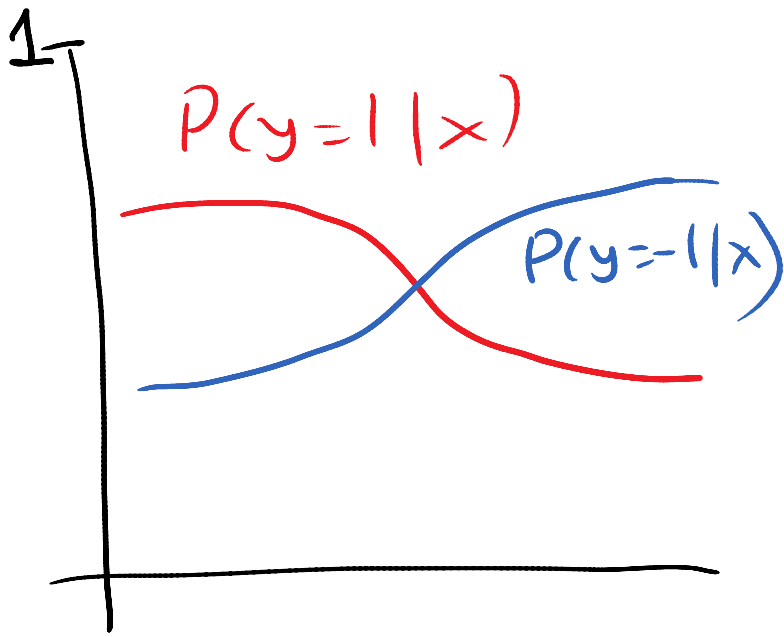
tween the two, but closer to Scenario 2. First we generated 10 means m_k from a bivariate Gaussian distribution $N((1,0)^T, \mathbf{I})$ and labeled this class BLUE. Similarly, 10 more were drawn from $N((0,1)^T, \mathbf{I})$ and labeled class ORANGE. Then for each class we generated 100 observations as follows: for each observation, we picked an m_k at random with probability $1/10$, and then generated a $N(m_k, \mathbf{I}/5)$, thus leading to a mixture of Gaussian clusters for each class. Figure 2.4 shows the results of classifying 10,000 new observations generated from the model. We compare the results for least squares and those for k-nearest neighbors for a range of values of k.

Good classifiers vs. Bad classifiers

- Good classifiers have an expected loss that is closer to the Bayes Risk of the classification problem, given a certain choice of features.
- In practice, we do not know the optimum classifier, nor do we know the Bayes risk. We must make do with what we can measure on our data sample.

Good features vs. Bad features





A Fairy Tale

Once upon a time there were two neighboring farmers, Jed and Ned. Each owned a horse, and the horses both liked to jump the fence between the two farms. Clearly, the farmers needed some means to tell whose horse was whose.

So Jed and Ned got together and agreed on a scheme for discriminating between the horses. Jed would cut a small notch in one ear of his horse. Not a big, painful notch, but one just big enough to be seen. Well, wouldn't you know it, the day after Jed cut the notch in his horse's ear, Ned's horse got caught on the barbed wire fence and tore his ear in exactly the same way.

A Fairy Tale (contd.)

Something else had to be devised, so Ned tied a big blue bow on the tail of his horse. But the next day, Jed's horse jumped the fence, ran into the field where Ned's horse was grazing and chewed the bow right off the horse's tail. Ate the whole bow!

Finally, Jed suggested, and Ned concurred, that they should pick a feature that was less apt to change. Height seemed like a good feature to use. But were the heights different? Well, each farmer went and measured his horse, and do you know what? The brown horse was a full two inches taller than the white one!

A Fairy Tale (contd.)

Something else had to be devised, so Ned tied a big blue bow on the tail of his horse. But the next day, Jed's horse jumped the fence, ran into the field where Ned's horse was grazing and chewed the bow right off the horse's tail. Ate the whole bow!

Finally, Jed suggested, and Ned concurred, that they should pick a feature that was less apt to change. Height seemed like a good feature to use. But were the heights different? Well, each farmer went and measured his horse, and do you know what? The brown horse was a full two inches taller than the white one!

Moral

- When you have difficulty in classification do not just look for esoteric mathematical tricks; instead find better measurements (features).
- Machine learning people need to talk to the domain experts.

- [Source of the fairy tale: Chapter 14, B.K.P. Horn, *Robot Vision*]

ROC curves & Precision-Recall curves

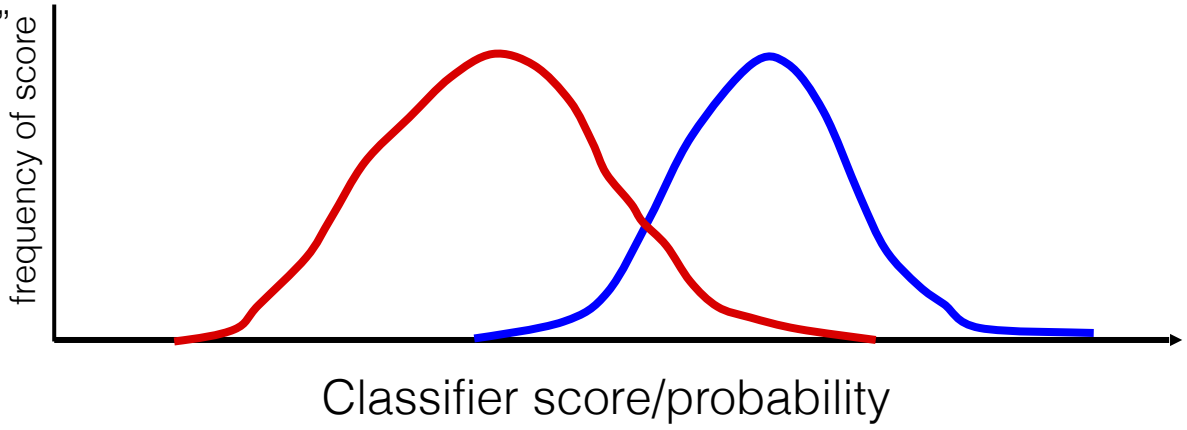
- Bayes-optimal classifiers and Bayes risk are theoretical concepts. Not available in practice.
- What we do in practice is to train a neural network (or some other classifier/regressor) on a data set. Typically this tries to model the posterior probability.
- Often the probability is not well-calibrated, so we need to pay more attention to the threshold-setting process. This leads us to ROC and precision-recall curves.

Defining false positives, false negative, etc.

[We will consider only binary classifiers for now]

Distribution of classifier "scores" of healthy and unhealthy individuals in a test set

(score could be a probability, but need not be)



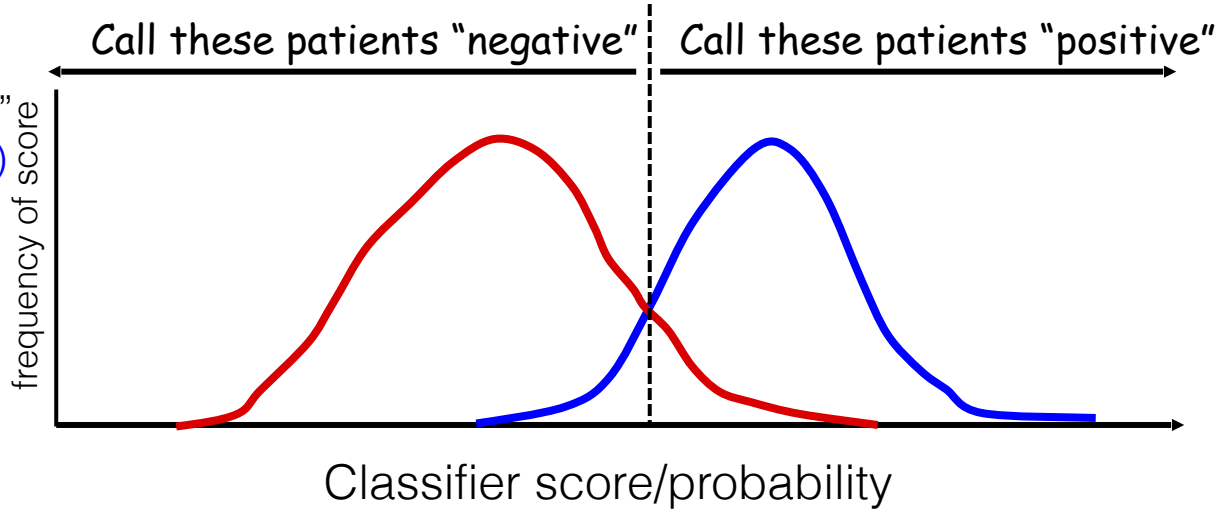
[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Defining false positives, false negative, etc.

Choose a threshold on the score/probabilistic output, and call all samples above it a "1" (e.g. "unhealthy) and all those below it a "-1" (e.g. "healthy").

Distribution of classifier "scores" of healthy (-1) and unhealthy (1) individuals in a test set.

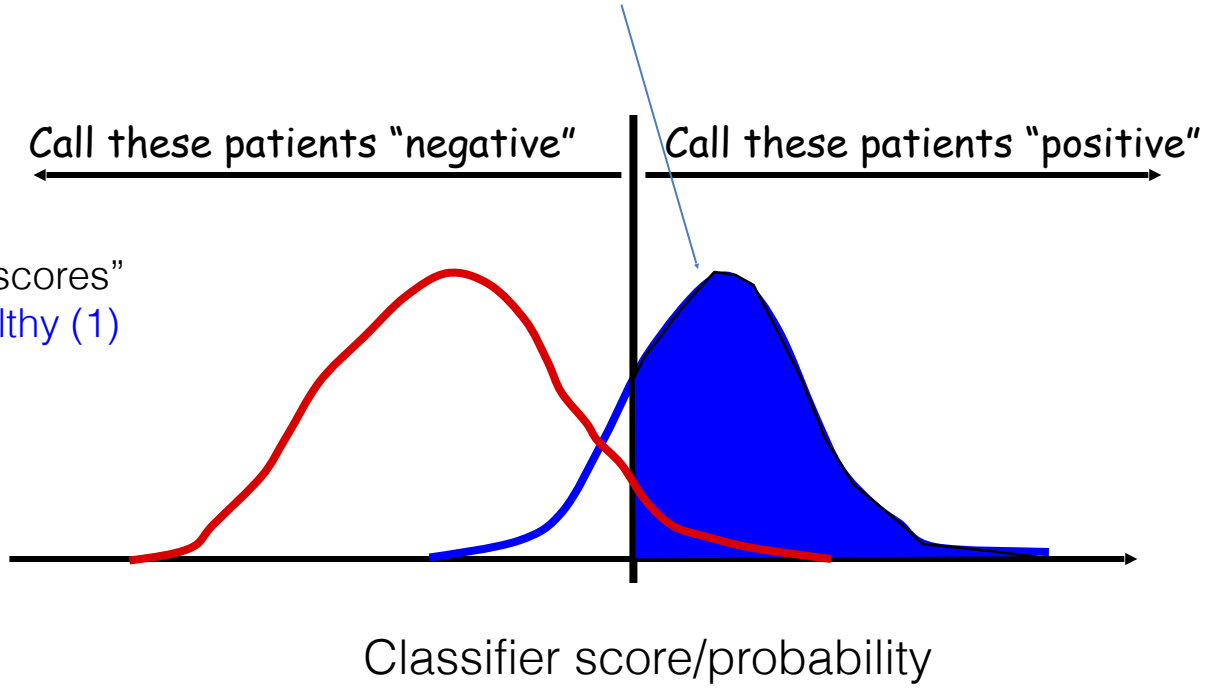
(score could be a probability, but need not be)



[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Definitions: True Positives (TP)

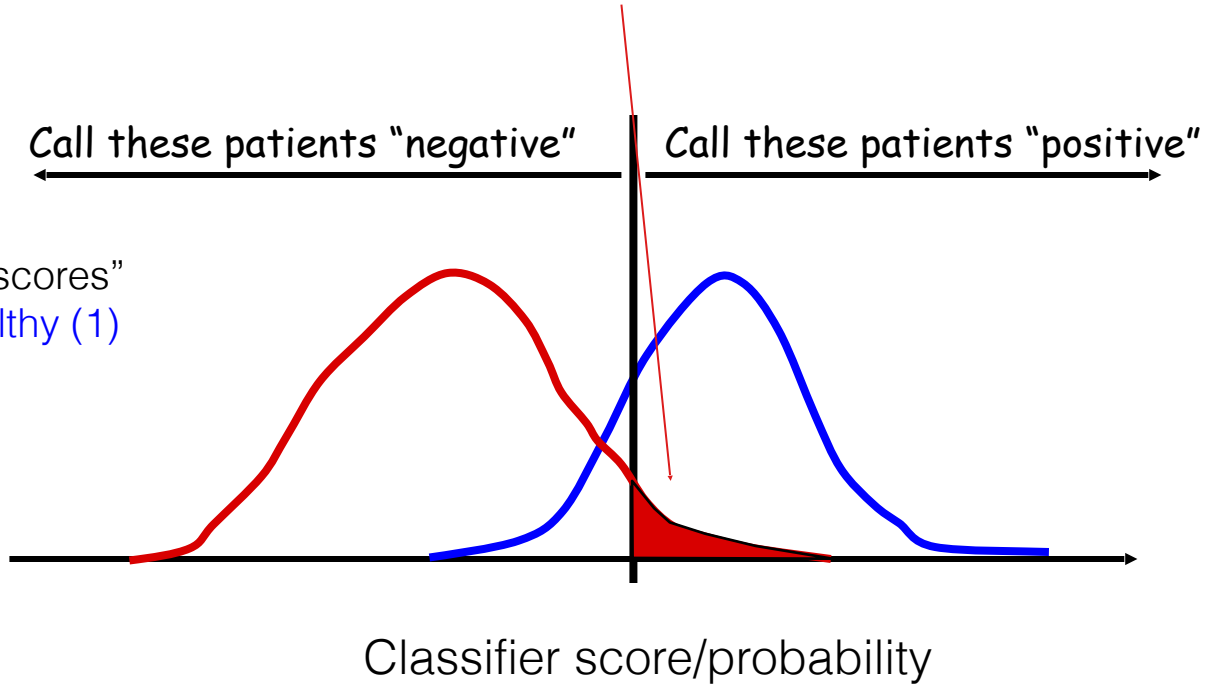
Distribution of classifier "scores" of healthy (-1) and unhealthy (1) individuals in a test set.



[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Definitions: False Positives (FP)

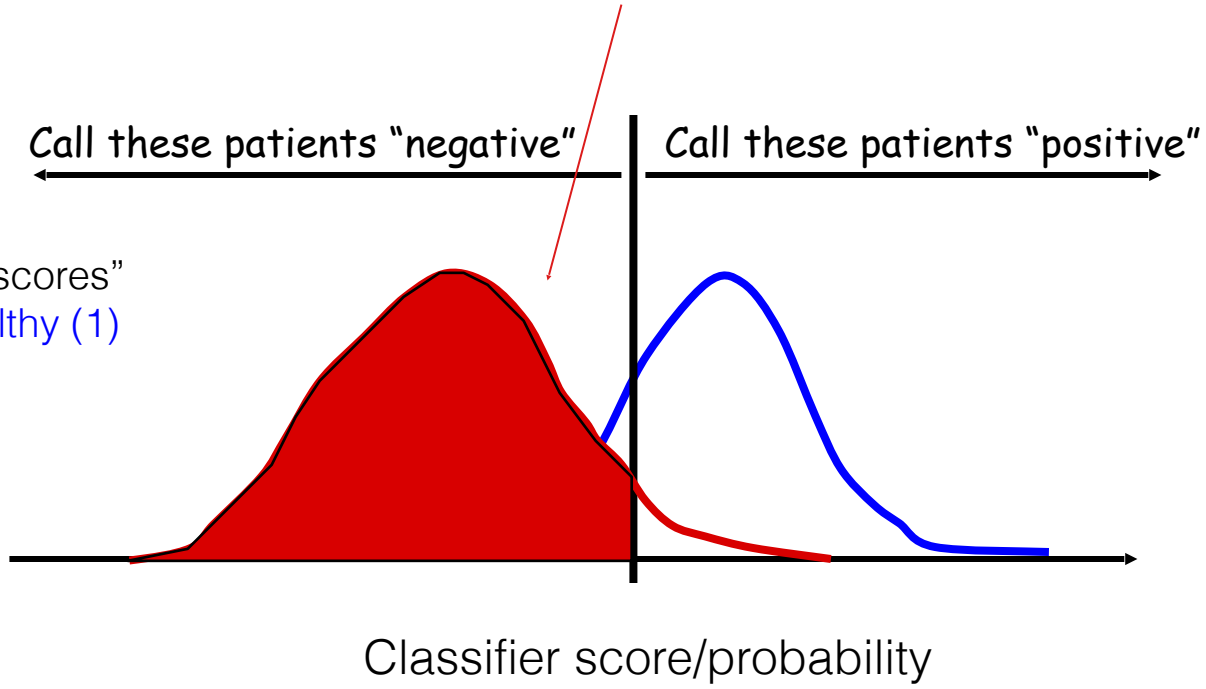
Distribution of classifier "scores" of healthy (-1) and unhealthy (1) individuals in a test set.



[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Definitions: True Negatives (TN)

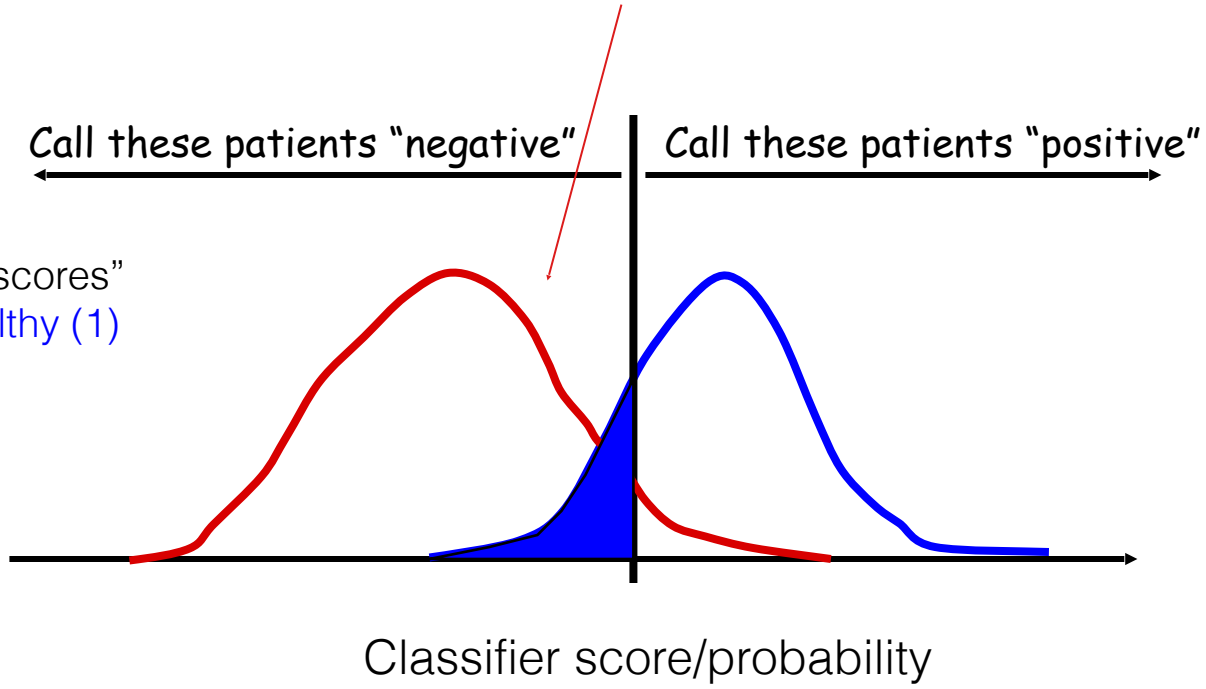
Distribution of classifier "scores" of healthy (-1) and unhealthy (1) individuals in a test set.



[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

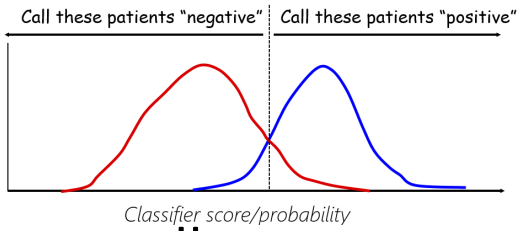
Definitions: False Negatives (FN)

Distribution of classifier "scores" of healthy (-1) and unhealthy (1) individuals in a test set.



[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Summary of classifier decision outco



Once we set a *decision* threshold on the predictive score, all test data points fall into one of these four categories:

1. False Positive (FP)—person is truly a "-1" but called "1"
2. False Negative (FN)—person is truly a "1" but called "-1"
3. True Positive (TP) —person is truly a "1" and called "1"
4. True Negative (TN) —person is truly a "-1" and called "-1"

- Thus if N is total # of test points, then N=FP+FN+TP+TN
- FP and FN are mistakes when using the classifier.
- TP and TN are correct decisions when using the classifier.

Summary of classifier decision outcomes

Often we see these reported in the form of a *confusion matrix:*

		MODEL PREDICTIONS		
		Negative	*Positive*	
GROUND TRUTH	*Negative*	TN	FP	#actual negatives=TN+FP
	Positive	FN	TP	#actual positives=FN+TP

[Adapted https://hiplab.mc.vanderbilt.edu/people/malin/presentations/ROC_Curves.ppt]

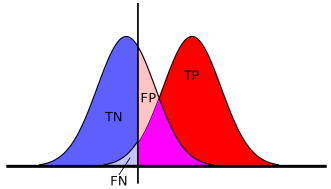
Summary of classifier decision outcomes

Rates: "normalize" by #samples who could have had that call:

- TP rate, TPR=TP/#actual positives=TP/(FN+TP), aka "Sensitivity"
- TN rate, TNR=TN/#actual negatives=TN/(TN+FP), aka "Specificity"
- FN rate, FNR=FN/#actual positives=1-TPR aka "Miss Rate"
- FP rate, FPR=FP/#actual negatives=1-TNR aka "Fall out"

		MODEL PREDICTIONS		
		Negative	*Positive*	
GROUND TRUTH	*Negative*	TN	FP	#actual negatives=TN+FP
	Positive	FN	TP	#actual positives=FN+TP

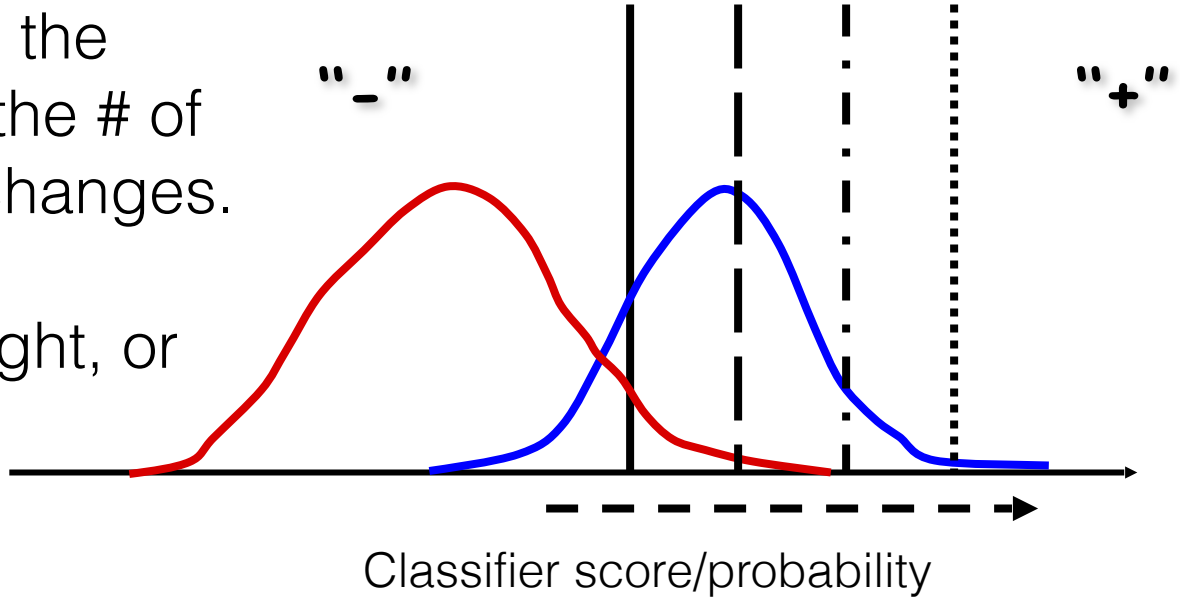
Back to our decision threshold



TP	FP
FN	TN

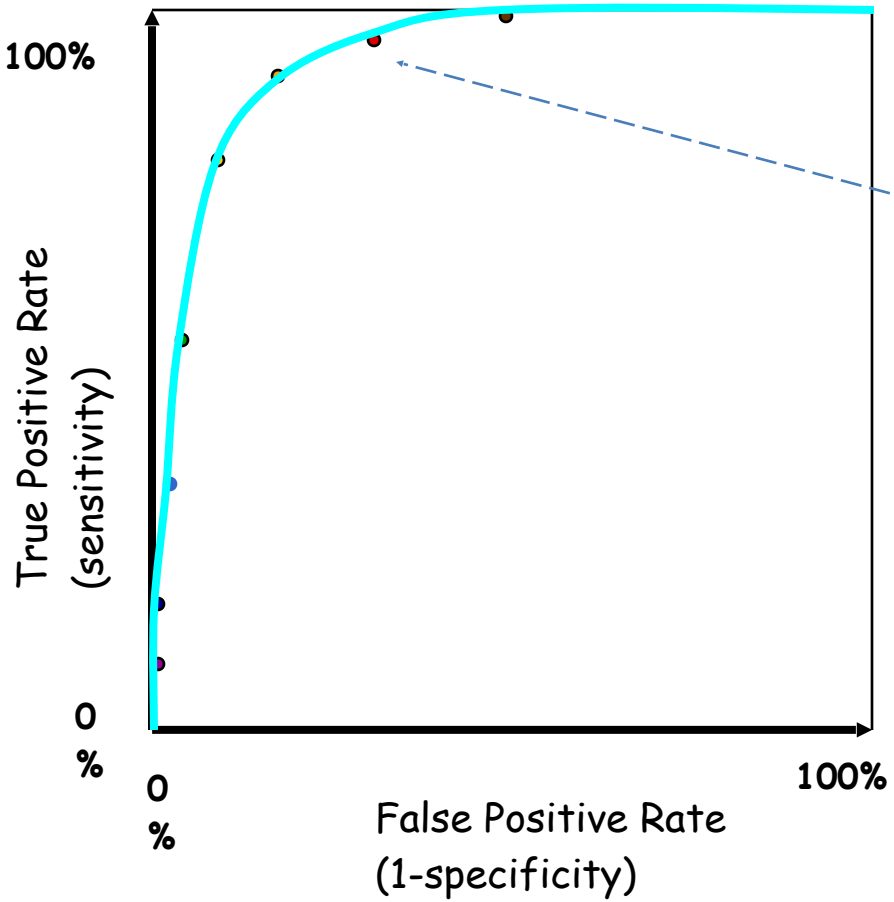
Every time we move the decision threshold, the # of FP, FN, TP and TN changes.

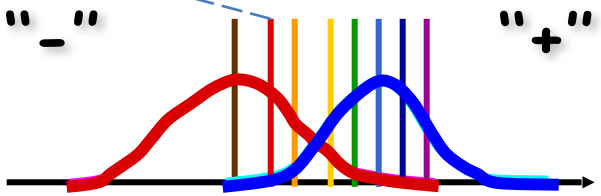
e.g. shifting to the right, or left



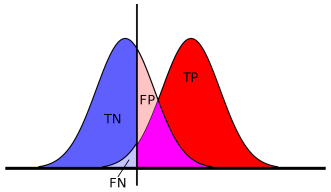
[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt], https://en.wikipedia.org/wiki/Receiver_operating_characteristic

As we shift it, we draw out an *ROC curve*





predictive distribution of classifier scores

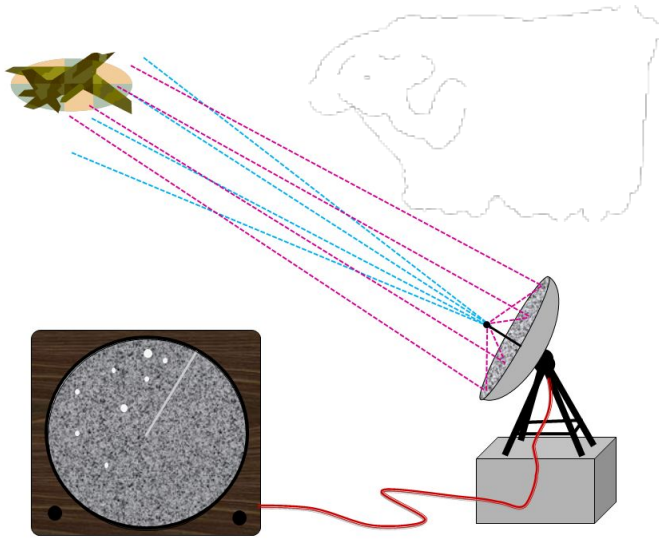


TP	FP
FN	TN

[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt], https://en.wikipedia.org/wiki/Receiver_operating_characteristic

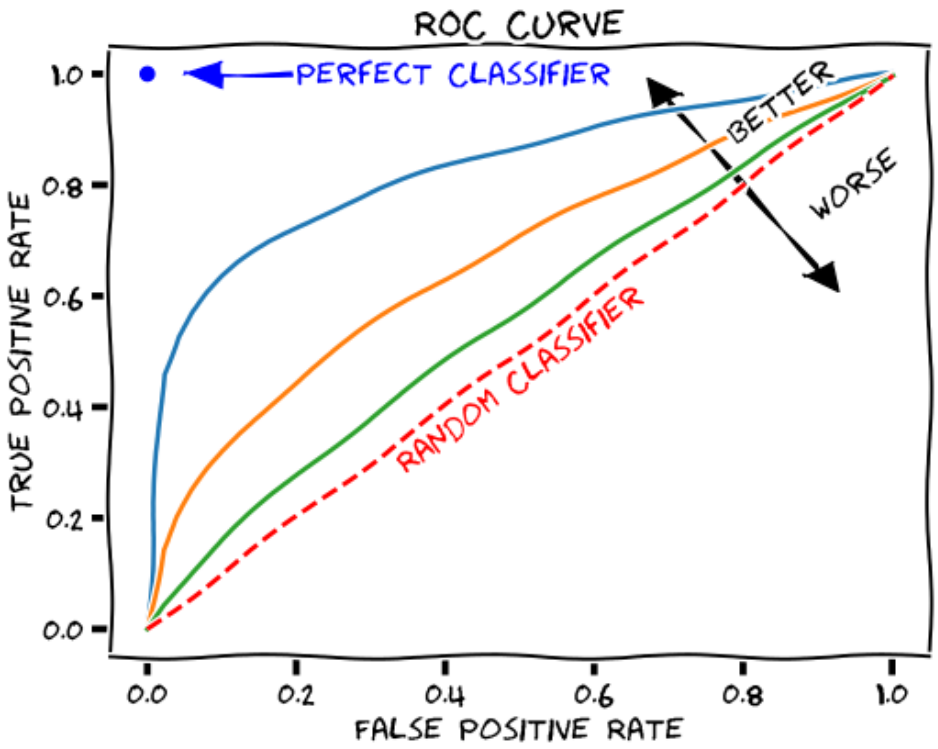
History of ROC curves

- “The ROC curve was first developed by electrical engineers and radar engineers during World War II (1939-45) for detecting enemy objects in battle fields and was soon introduced to psychology to account for perceptual detection of stimuli.”
- “ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.”



from http://en.wikipedia.org/wiki/Receiver_operating_characteristic

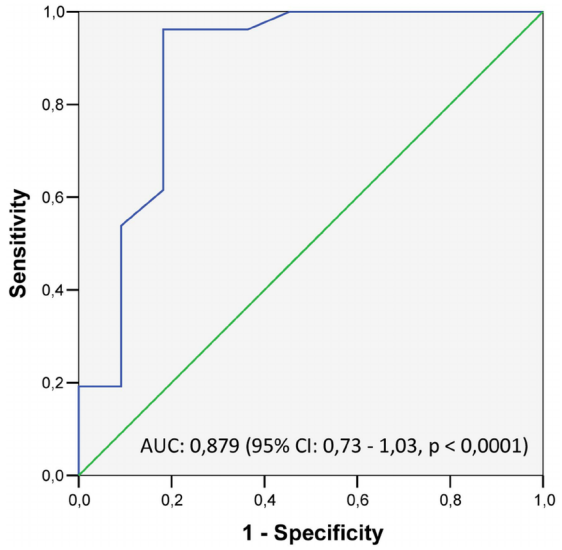
Comparing ROC curves across classifiers

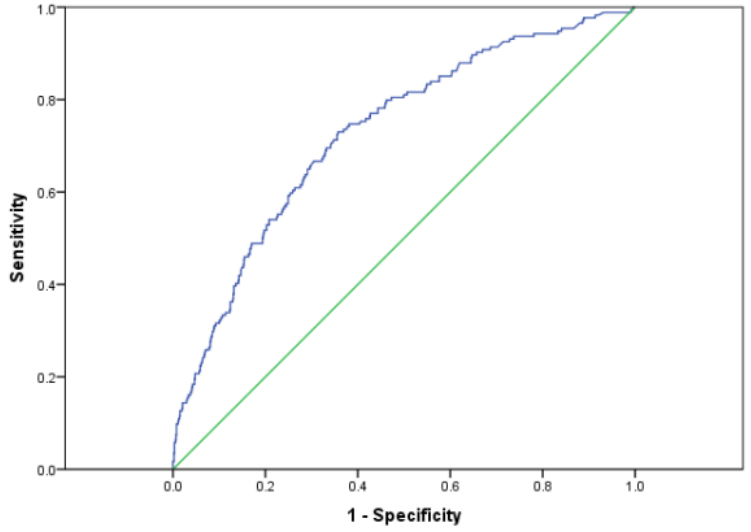


https://en.wikipedia.org/wiki/File:Roc-draft-xkcd-style.svg

An algorithm for making an ROC curve

A consequence of this algorithm is that the smoothness of the ROC curve is dependent on the # of points in it, is restricted by number of test points (and uniqueness of scores):

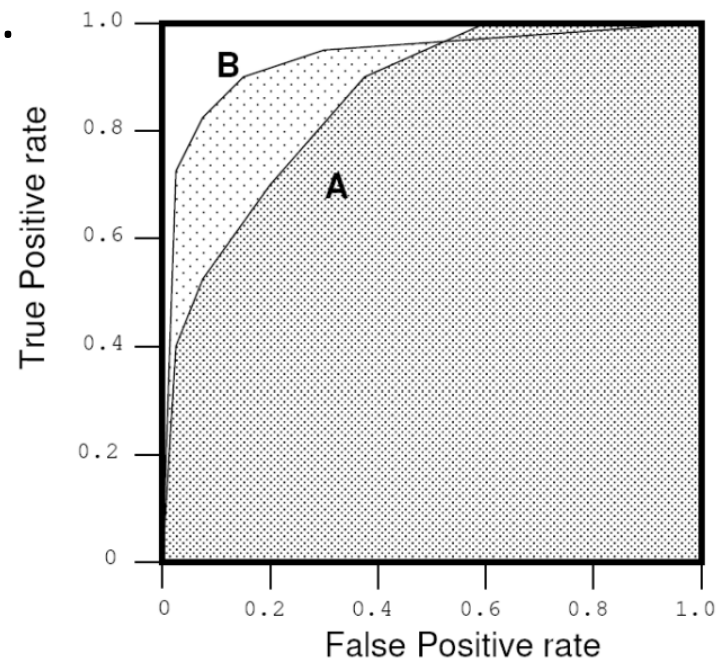




https://www.theanalysisfactor.com/what-is-an-roc-curve, https://www.researchgate.net/figure/The-Receiver-Operating-Characteristic-ROC-curve-of-different-OUES-cut-off-points_fig2_263971330

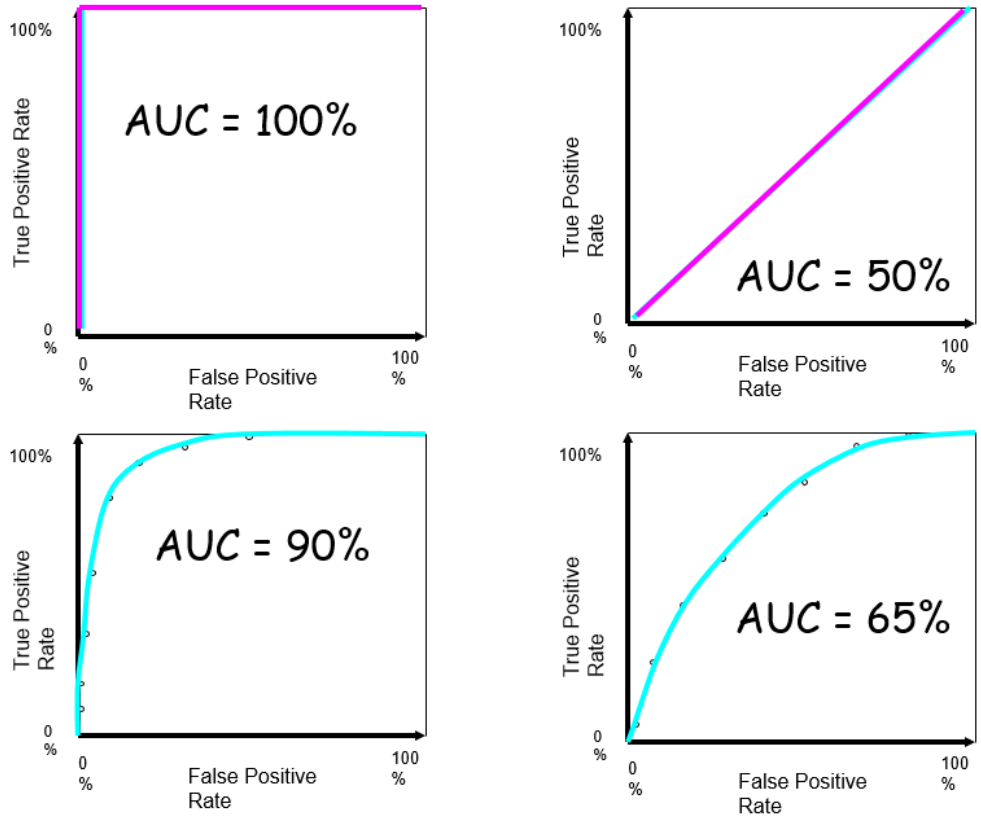
Summarizing ROCs with the Area Under the Curve (AUC)

- AUC: often used to compare classifiers.
- The bigger the AUC the better.
- AUC can be computed by a slight modification to the algorithm for constructing ROC curves—basically a simple form of integration to compute the area under the curve.



https://web.uvic.ca/~maryam/DMSpring94/Slides/9_roc.pdf

Summarizing ROCs with the Area Under the Curve (AUC)

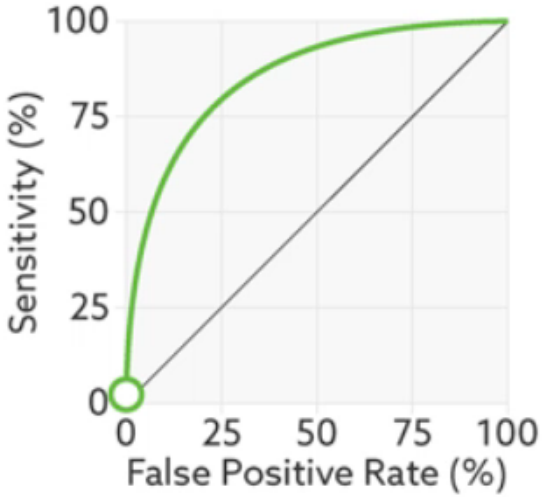


The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample.

[Adapted from http://www.lausanne.isb-sib.ch/~darlene/ms/SIB-ROC.ppt]

Visualization of score distributions wrt ROCs & AUC

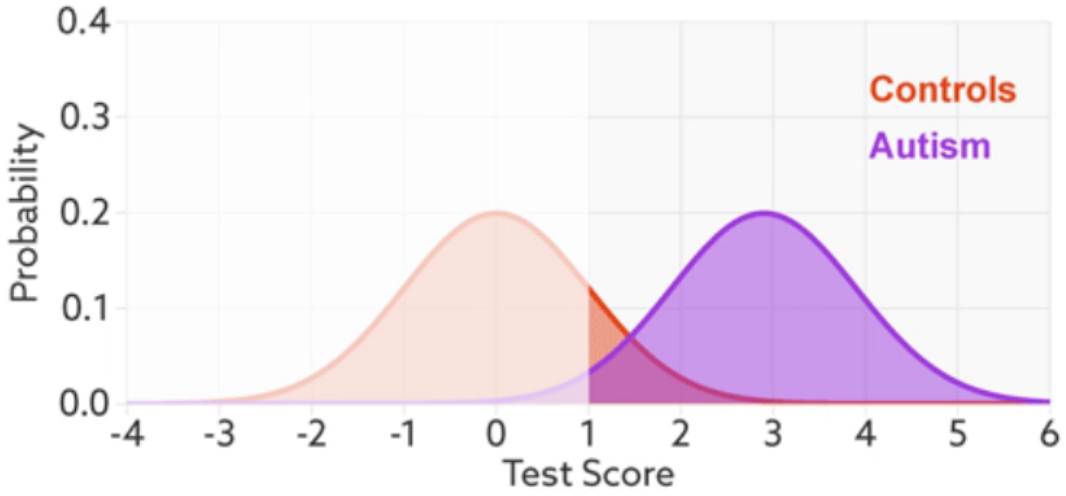
1. Sweeping a threshold through the predictions for one classifier traces out the ROC curve

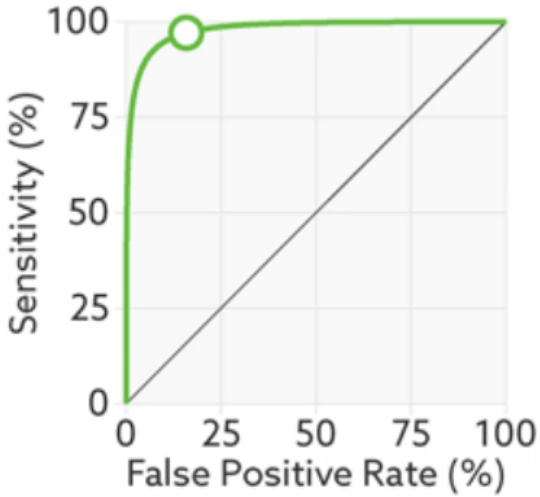


https://www.spectrumnews.org/opinion/viewpoint/quest-autism-biomarkers-faces-steep-statistical-challenges/

Visualization of score distributions wrt ROCs & AUC

2. The more separated the distribution of scores between the two classes, the larger the AUC (i.e. this is a continuum of different classifiers).





https://www.spectrumnews.org/opinion/viewpoint/quest-autism-biomarkers-faces-steep-statistical-challenges/

More on ROC curves

- ROC curves are insensitive to the balance of classes in the test set (because FPR and FNR are insensitive quantities).
- To compute the classification accuracy from an ROC we need to know the ratio of # actual positives to # actual negatives in the test set.
- Knowing this we can find a point on the graph with optimal classification accuracy.
- Sometimes people use Precision-Recall curves instead of ROC because they are sensitive to the balance of classes in the test set.

Summary of classifier decision outcomes

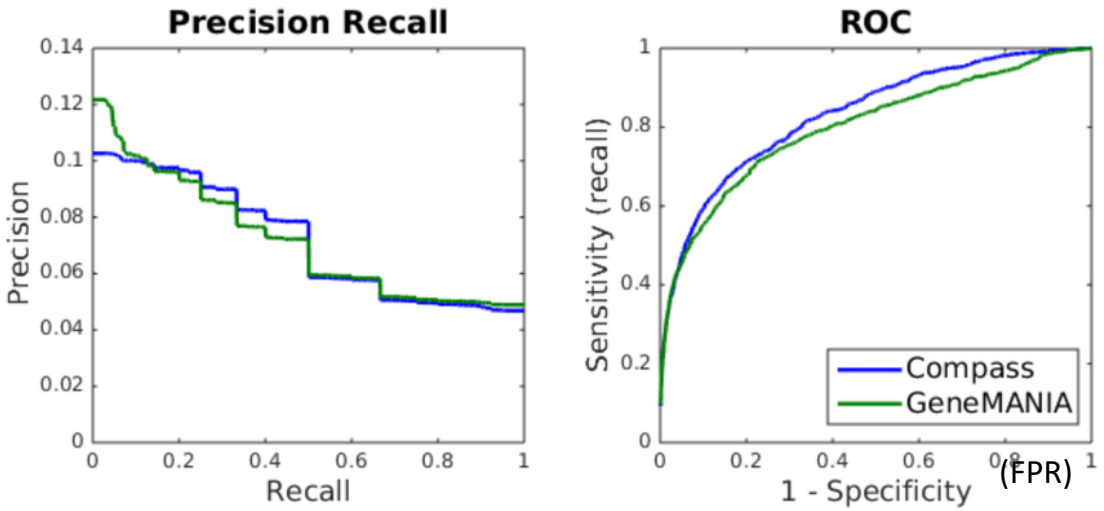
Rates: "normalize" by #samples who could have had that call:

- TP rate, TPR=TP/#actual positives=TP/(FN+TP), aka "Sensitivity"
- TN rate, TNR=TN/#actual negatives=TN/(TN+FP), aka "Specificity"
- FN rate, FNR=FN/#actual positives=1-TPR aka "Miss Rate"
- FP rate, FPR=FP/#actual negatives=1-TNR aka "Fall out"
- Precision=TP/(#predicted positive)=TP/(TP+FP)—this now depends on class balance in test set.
- Recall=TP/(#actual positives)=TPR

		MODEL PREDICTIONS		
		Negative	*Positive*	
GROUND TRUTH	*Negative*	TN	FP	#actual negatives=TN+FP
	Positive	FN	TP	#actual positives=FN+TP

Precision Recall Curves

- Can draw analogous plots to ROC, but now with Precision on the vertical axis and Recall (TPR) on the horizontal axis.
- Wrt ROC: replace the FPR with the Precision, and flip the axes:

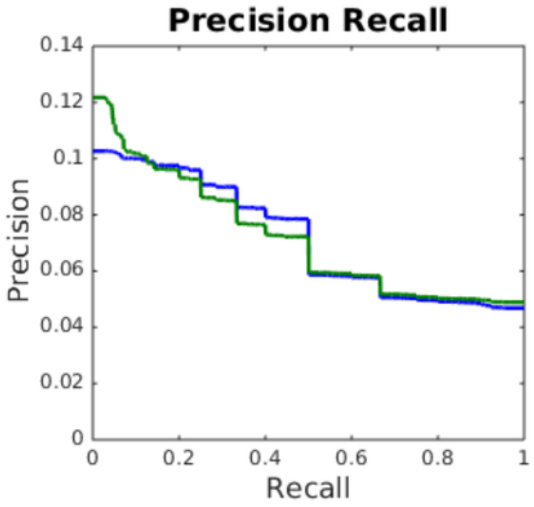


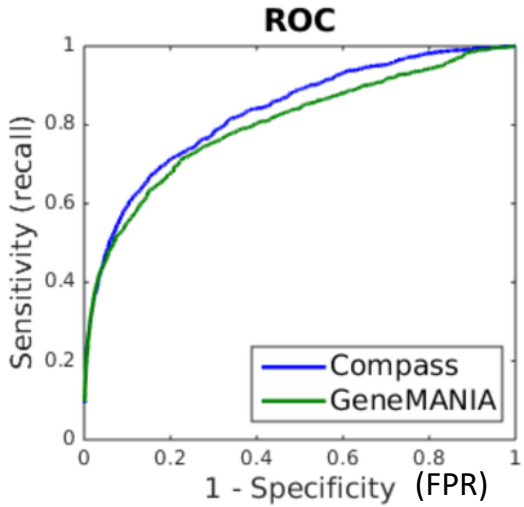
https://www.researchgate.net/figure/Average-precision-recall-and-ROC-curves-for-the-GO-rollback-benchmark-The-curves_fig2_281172305

Precision Recall Curves

- Can draw analogous plots to ROC, but now with Precision on the vertical axis and Recall (TPR) on the horizontal axis.
- Wrt ROC: replace the FPR with the Precision, and flip the axes:

- ROC curves useful when we want invariance to class distribution.
- Precision-recall curves useful when care about (and know) the balance of the classes at test time.

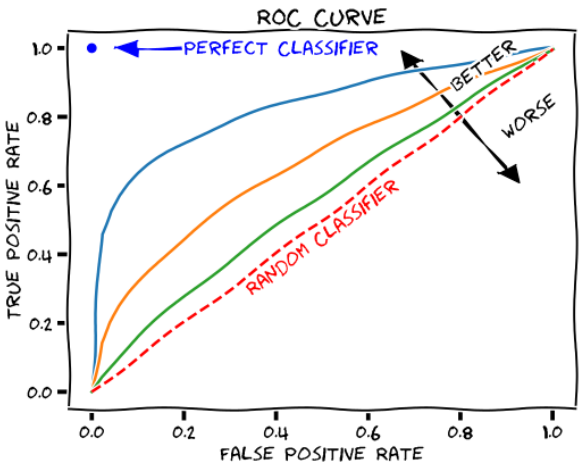




https://www.researchgate.net/figure/Average-precision-recall-and-ROC-curves-for-the-GO-rollback-benchmark-The-curves_fig2_281172305

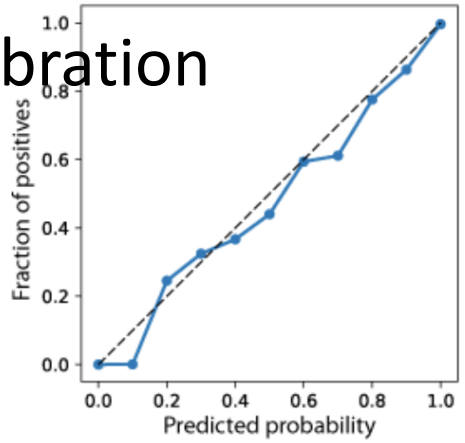
Summary of ROC curves and their utility

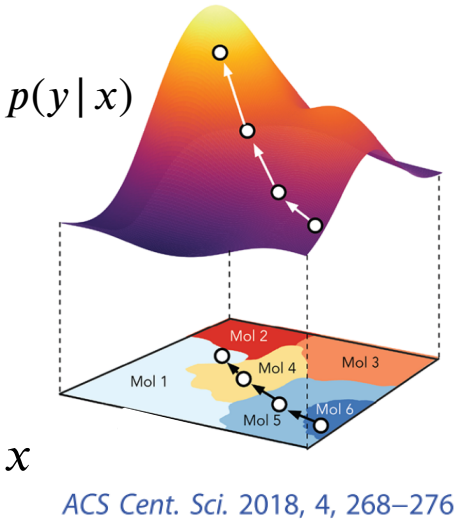
- Gives more nuanced understanding than counting the # of misclassifications.
- Does not require a decision threshold.
- Summarizes performance of binary classification models, across all possible trade-offs in decision making FP and FN.
- Does not care about model calibration (can be a pro or a con).
- Can compare classifiers by comparing their AUC summary, or using the entire ROC curves.



If you care about (probabilistic) model calibration

- Then you should NOT use ROC curves to evaluate, because they don't care about "calibrated uncertainty" of the predictions.
- Calibrated uncertainty can be very important in medical applications, such as to determine treatments, or administering invasive diagnostics.
- Calibration also come into ML-based design---e.g. in small molecule engineering (e.g. use a predictive model to design the best binder to drug target).
- Also in "active learning".





ACS Cent. Sci. 2018, 4, 268–276

Many other evaluations—dictated by the domain of application

- Predict a Ranking (of webpages)
  - Users only look at top 4
  - Sort by f(x|w,b)
- Precision @4 =1/2
  - Fraction of top 4 relevant
- Recall @4 =2/3
  - Fraction of relevant in top 4
- Top of Ranking Only!

Image Source: http://pmtk3.googlecode.com/svn-history/r785/trunk/docs/demos/Decision_theory/PRhand.html