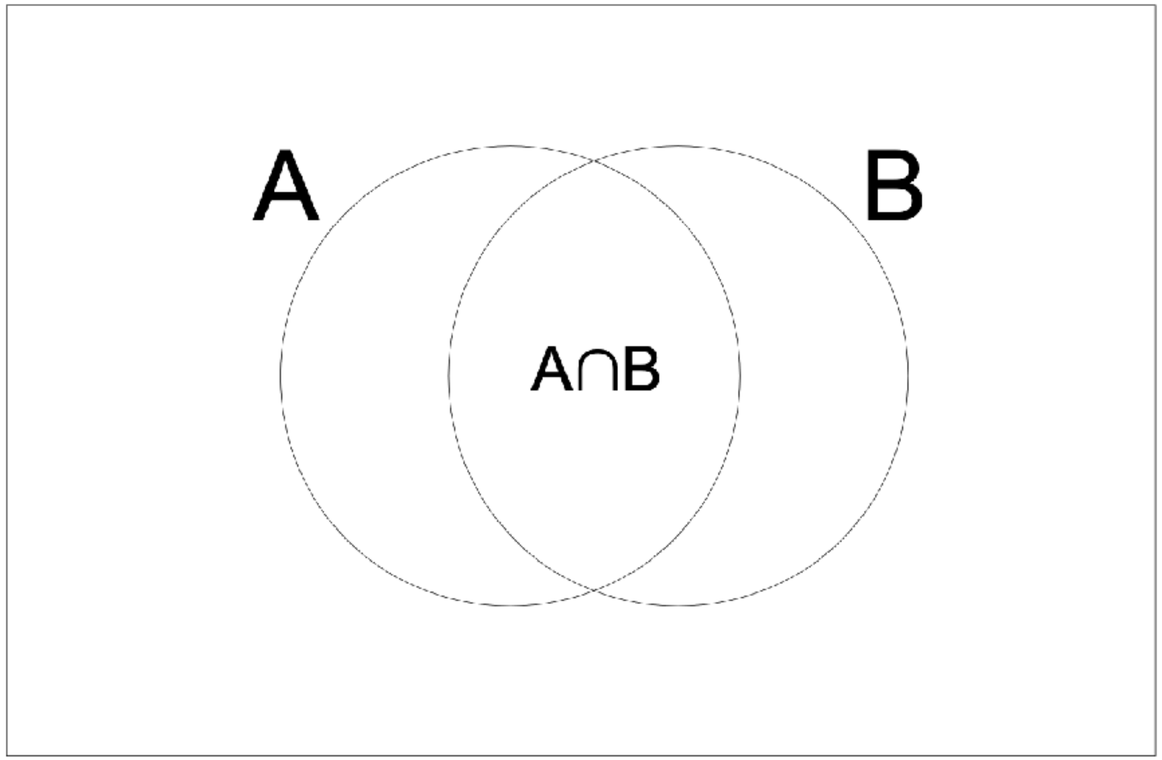
Lecture 10 pre-video

Error metrics



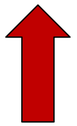
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Rev Thomas Bayes

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Error measures and metrics

	predicted+	predicted-
true label+	true label+ and predicted+	true label+ and predicted-
true label-	true label- and predicted+	true label- and predicted-



larger preferred

True positive rate: P(predicted + | true label +)

$= sensitivity = recall$

smaller preferred

False positive rate: P(predicted+ | true label -)

larger preferred

True negative rate: P(predicted - | true label-)

$= specificity$

smaller preferred

False negative rate: P(predicted- | true label +)

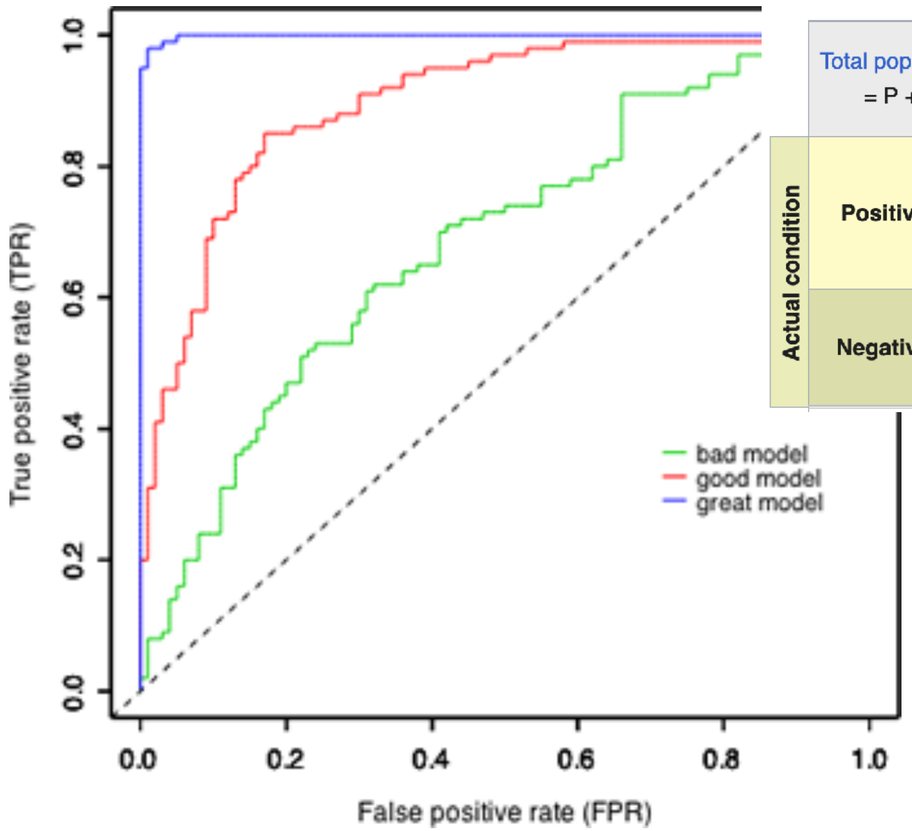
Error Metrics and Evaluation

		Predicted condition		Sources: [6][7][8][9][10][11][12][13][14] view·talk·edit	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

https://en.wikipedia.org/wiki/Precision_and_recall

Receiver Operating Characteristic

ROC-AUC

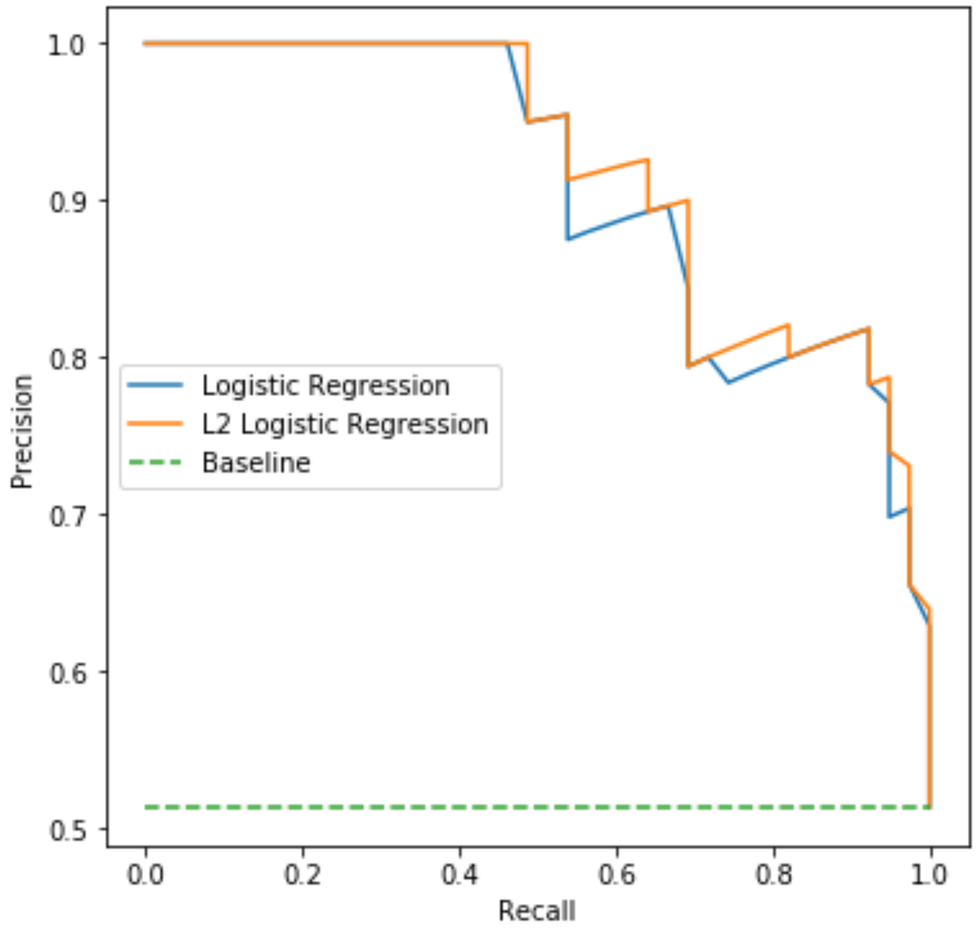


		Predicted condition				
	Total population = P + N	Positive (PP)	Negative (PN)			
Actual condition	Positive (P)	**True positive** (TP), hit	**False negative** (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$	
	Negative (N)	**False positive** (FP), type I error, false alarm, overestimation	**True negative** (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$	

The value can range from 0 to 1. However AUC score of a random classifier for balanced data is 0.5

Courtesy of Alvira Swalin

Precision - Recall curves



		Predicted condition		
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	

Error metrics

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Slides in this presentation are from material kindly provided by Zhuowen Tu

- HOW GOOD?
  - Can use the loss function
    - but may want other views of performance
    - comparing models with different loss functions
  - Many metrics
    - applicable to only one kind of problem
    - each metric a different aspect of performance

Scoring	Function	Comment
Classification		
'accuracy'	metrics.accuracy_score	
'balanced_accuracy'	metrics.balanced_accuracy_score	
'average_precision'	metrics.average_precision_score	
'neg_brier_score'	metrics.brier_score_loss	
'f1'	metrics.f1_score	for binary targets
'f1_micro'	metrics.f1_score	micro-averaged
'f1_macro'	metrics.f1_score	macro-averaged
'f1_weighted'	metrics.f1_score	weighted average
'f1_samples'	metrics.f1_score	by multilabel sample
'neg_log_loss'	metrics.log_loss	requires `predict_proba` support
'precision' etc.	metrics.precision_score	suffixes apply as with 'f1'
'recall' etc.	metrics.recall_score	suffixes apply as with 'f1'
'jaccard' etc.	metrics.jaccard_score	suffixes apply as with 'f1'
'roc_auc'	metrics.roc_auc_score	
'roc_auc_ovr'	metrics.roc_auc_score	
'roc_auc_ovo'	metrics.roc_auc_score	
'roc_auc_ovr_weighted'	metrics.roc_auc_score	
'roc_auc_ovo_weighted'	metrics.roc_auc_score	
Clustering		
'adjusted_mutual_info_score'	metrics.adjusted_mutual_info_score	
'adjusted_rand_score'	metrics.adjusted_rand_score	
'completeness_score'	metrics.completeness_score	
'fowlkes_mallows_score'	metrics.fowlkes_mallows_score	
'homogeneity_score'	metrics.homogeneity_score	
'mutual_info_score'	metrics.mutual_info_score	
'normalized_mutual_info_score'	metrics.normalized_mutual_info_score	
'v_measure_score'	metrics.v_measure_score	
Regression		
'explained_variance'	metrics.explained_variance_score	
'max_error'	metrics.max_error	
'neg_mean_absolute_error'	metrics.mean_absolute_error	
'neg_mean_squared_error'	metrics.mean_squared_error	
'neg_root_mean_squared_error'	metrics.mean_squared_error	
'neg_mean_squared_log_error'	metrics.mean_squared_log_error	
'neg_median_absolute_error'	metrics.median_absolute_error	
'r2'	metrics.r2_score	
'neg_mean_poisson_deviance'	metrics.mean_poisson_deviance	
'neg_mean_gamma_deviance'	metrics.mean_gamma_deviance	

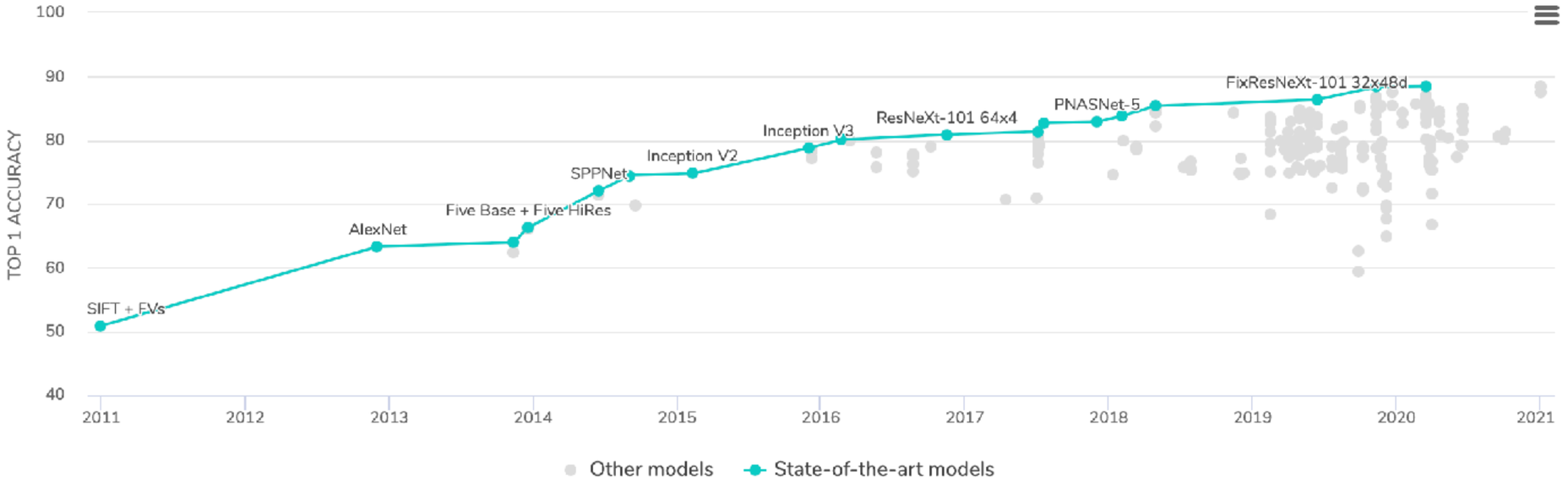
Error Metrics and Loss Functions

- One thing that separates modern machine learning from the efforts in traditional AI is the establishment of benchmarks under widely accepted common evaluation metrics.

- Being able to faithfully compare the performances of different machine learning algorithms/systems significantly propel the advancement of machine learning field.

- Do not confuse loss functions (error terms + regularization) that are optimized when training ML algorithms with a metric used to evaluate performance. Different algorithms use different loss functions. But to compare two algorithms we use the same metric on the same benchmark dataset.

Image Classification on ImageNet



View Top 1 Accuracy All models Edit

RANK	MODEL	TOP 1 ACCURACY	TOP 5 ACCURACY	NUMBER OF PARAMS	EXTRA TRAINING DATA	PAPER	CODE	RESULT	YEAR
1	**FixEfficientNet-L2**	88.5%	98.7%	480M	✓	Fixing the train-test resolution discrepancy: FixEfficientNet			2020
2	**NoisyStudent** (EfficientNet-L2)	88.4%	98.7%	480M	✓	Self-training with Noisy Student improves ImageNet classification			2019
3	**ViT-H/14**	88.36%		632M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at			2020

Images

ImageNet

Introduced by Jia Deng et al. in ImageNet: A large-scale hierarchical image database

The **ImageNet** dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection. The publicly released dataset contains a set of manually annotated training images. A set of test images is also released, with the manual annotations withheld. ILSVRC annotations fall into one of two categories: (1) image-level annotation of a binary label for the presence or absence of an object class in the image, e.g., "there are cars in this image" but "there are no tigers," and (2) object-level annotation of a tight bounding box and class label around an object instance in the image, e.g., "there is a screwdriver centered at position (20,25) with width of 50 pixels and height of 30 pixels". The ImageNet project does not own the copyright of the images, therefore only thumbnails and URLs of images are provided.

- Total number of non-empty WordNet synsets: 21841
- Total number of images: 14197122
- Number of images with bounding box annotations: 1,034,908
- Number of synsets with SIFT features: 1000
- Number of images with SIFT features: 1.2 million

Source: ImageNet Large Scale Visual Recognition Challenge

Summary of the problem

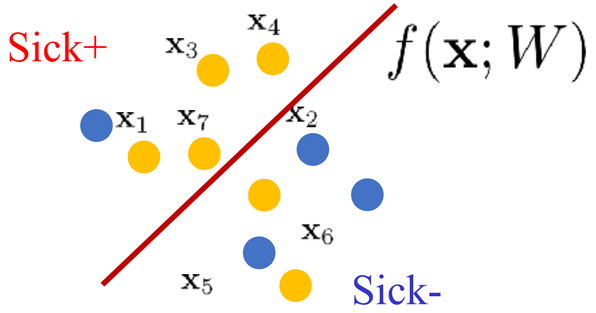
$$S_{training} = \{(\mathbf{x}_i, y_i), i = 1..n\} \qquad \mathbf{x} = (x_1, ..., x_m), x_i \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^m$$

$$y \in \{-1, +1\} \qquad y = -1\text{: sick-}$$

$$y = +1\text{: sick+}$$

Classifier: $Classify = f(\mathbf{x}; W) \in \{-1, +1\}$

Model parameter to be learned: W



Training error:

$$e_{training} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \neq f(\mathbf{x}_i; W))$$

$$e_{training} = \frac{3}{10} = 0.3$$

Testing error:

$$e_{testing} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{1}(y_i \neq f(\mathbf{x}_i; W))$$

$$e_{testing} = \frac{4}{11} = 0.3636$$

Just use the misclassification rate, right?

Classification error: $e = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \neq f(\mathbf{x}_i; W))$ AKA (1.0 - accuracy score)

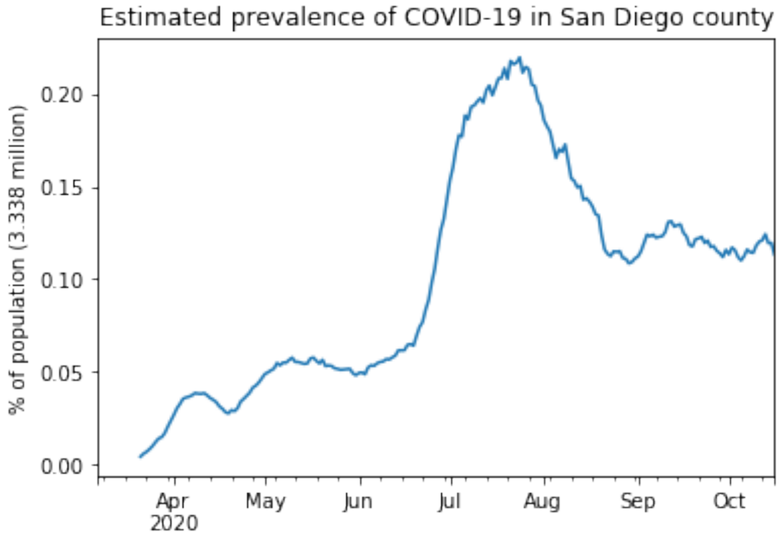
A sound metric when classes are balanced, but in practice datasets often aren't

much greater

Often: $\sum_{i=1}^{n} \mathbf{1}(y_i = 0)) >> \sum_{i=1}^{n} \mathbf{1}(y_i = 1))$

# of negatives | # of positives

For example, if we have **1,000** negative samples and 1 positive sample, we can trivially predicted every input sample as negative and do quite well!



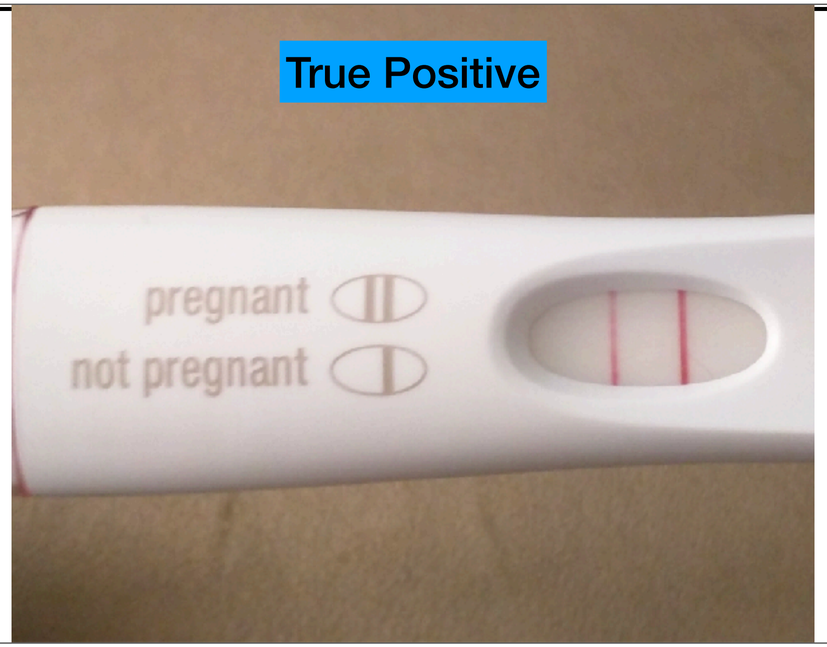
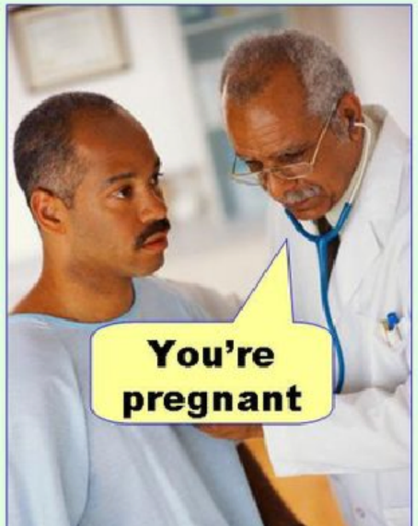
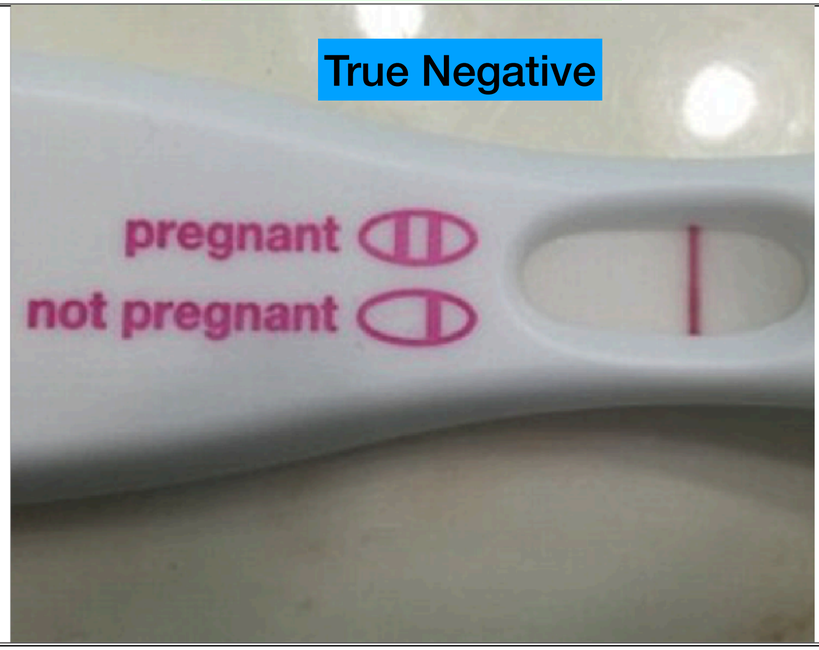
Classification error: $e = \frac{1}{1{,}001}$

Our classifier was 94% accurate at predicting whether a sample was malignant, benign, or non-cancerous

But our dataset was 72% benign, 16% non-cancerous, and 12% malignant. So what would our accuracy be if our classifier was just guessing “benign” without caring about the variables?

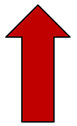
Our classifier was 94% accurate at predicting whether a sample was malignant, benign, or non-cancerous

But 36% of malignant cancers were misclassified as benign

	Predicted +	Predicted -
Truly +	True Positive 	**Type II error** (false negative) 
Truly -	**Type I error** (false positive) 	True Negative 

Error measures and metrics

	predicted+	predicted-
true label+	true label+ and predicted+	true label+ and predicted-
true label-	true label- and predicted+	true label- and predicted-



larger preferred

True positive rate: P(predicted + | true label +)

$= sensitivity = recall$

smaller preferred

False positive rate: P(predicted+ | true label -)

larger preferred

True negative rate: P(predicted - | true label-)

$= specificity$

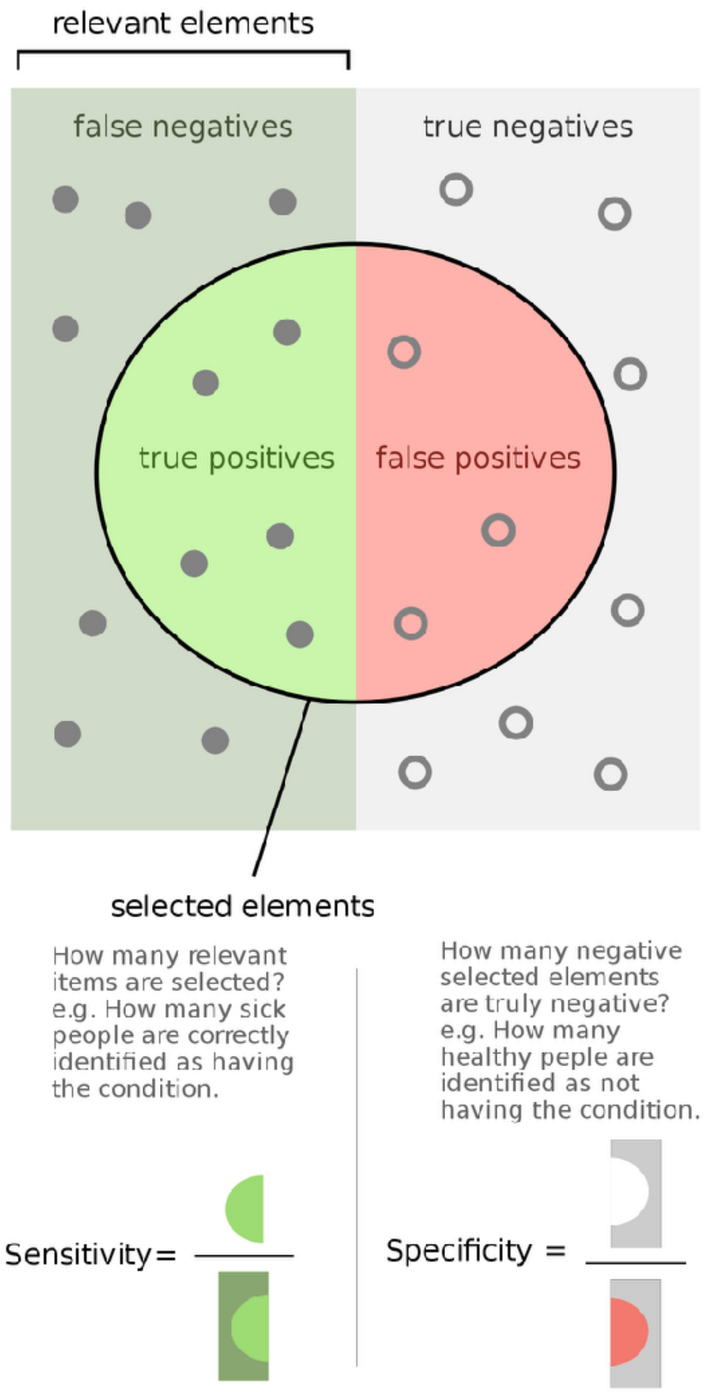
smaller preferred

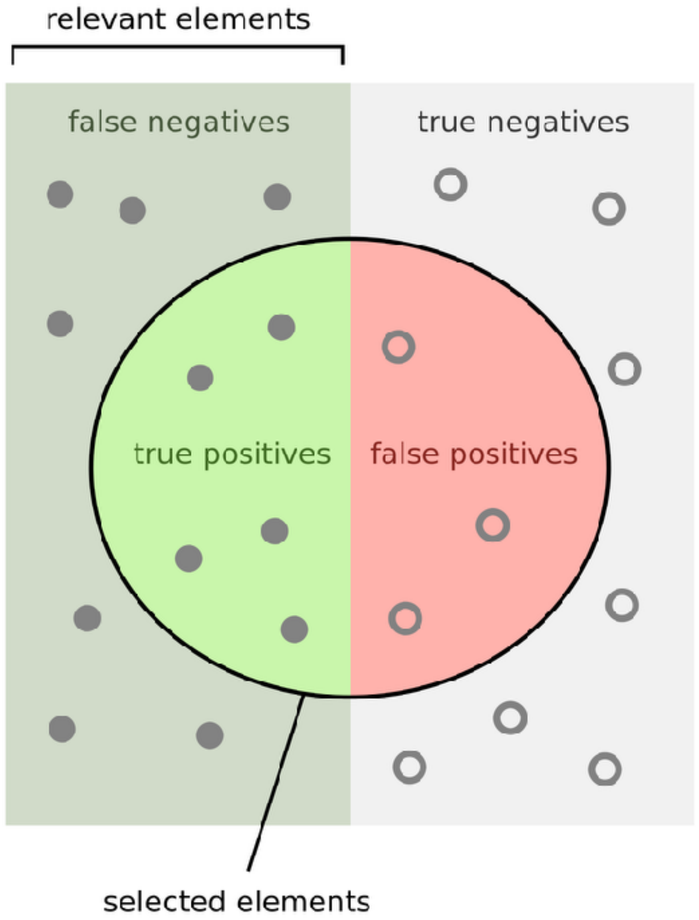
False negative rate: P(predicted- | true label +)

Error Metrics and Evaluation

		Predicted condition		Sources: [6][7][8][9][10][11][12][13][14] view·talk·edit	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

https://en.wikipedia.org/wiki/Precision_and_recall


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.
Sensitivity=
How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.
Specificity =

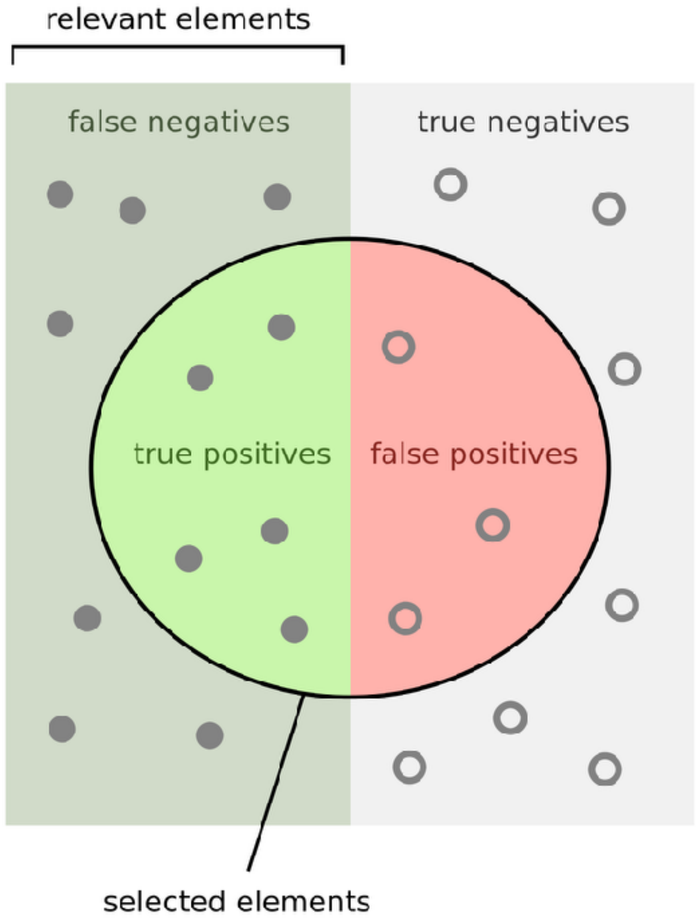


How many selected items are relevant?

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

How many relevant items are selected?

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

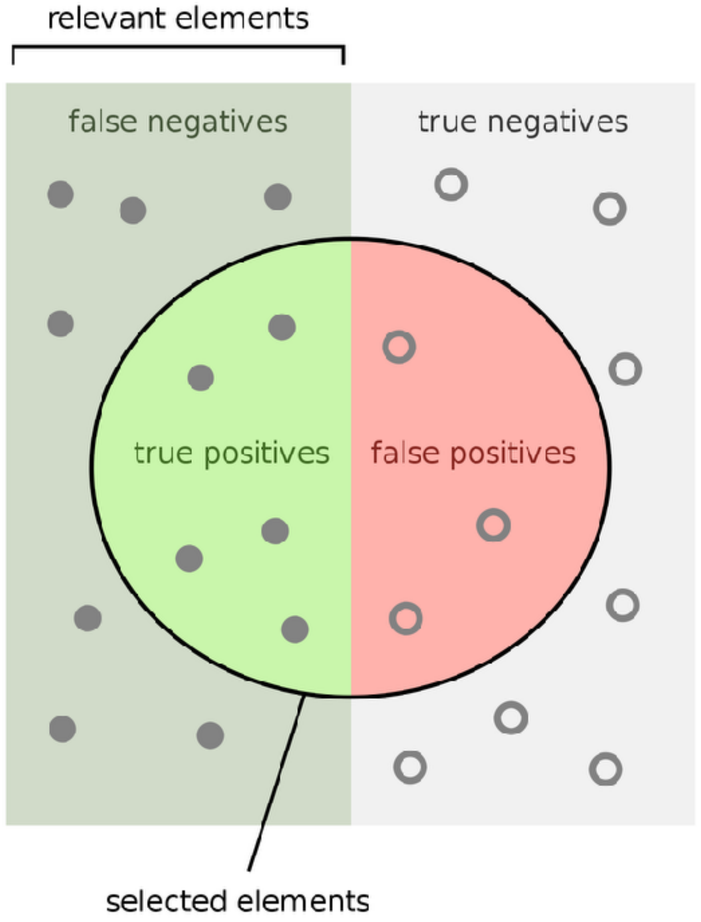


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$



How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

Intuition Test

Summary (confusion matrix)

y \ $f(\mathbf{x}; W)$	**Predicted +**	**Predicted -**	**Total**
Sick +	30	10	40
Sick -	10	50	60
Total	40	60	100

Sensitivity = TP / P

Specificity = TN / N

A. High sensitivity, low specificity

B. High sensitivity, high specificity

C. Low sensitivity, low specificity

D. Low sensitivity, high specificity

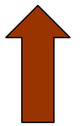
Error measures

Summary (confusion matrix)

y \ $f(\mathbf{x}; W)$	**Predicted +**	**Predicted -**	**Total**
Sick +	30	10	40
Sick -	10	50	60
Total	40	60	100

True positive rate: P(classify+ | sick +) $= \frac{30}{40} = 0.75$

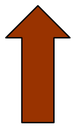
$= sensitivity = recall$



False positive rate: P(classify + | sick -) $= \frac{10}{60} = 0.167$

True negative rate: P(classify - | sick-) $= \frac{50}{60} = 0.833$

$= specificity$



False negative rate: P(classify - | sick +) $= \frac{10}{40} = 0.25$

Compare models

	Predicted +	Predicted -	Total
Sick +	30	10	40
Sick -	10	50	60
Total	40	60	100

TPR = 75%
FPR = 17%

	Predicted +	Predicted -	Total
Sick +	25	15	40
Sick -	1	59	60
Total	26	74	100

TPR = 63%
FPR = 2%

Compare models

	Predicted +	Predicted -	Total
Sick +	30	10	40
Sick -	10	50	60
Total	40	60	100

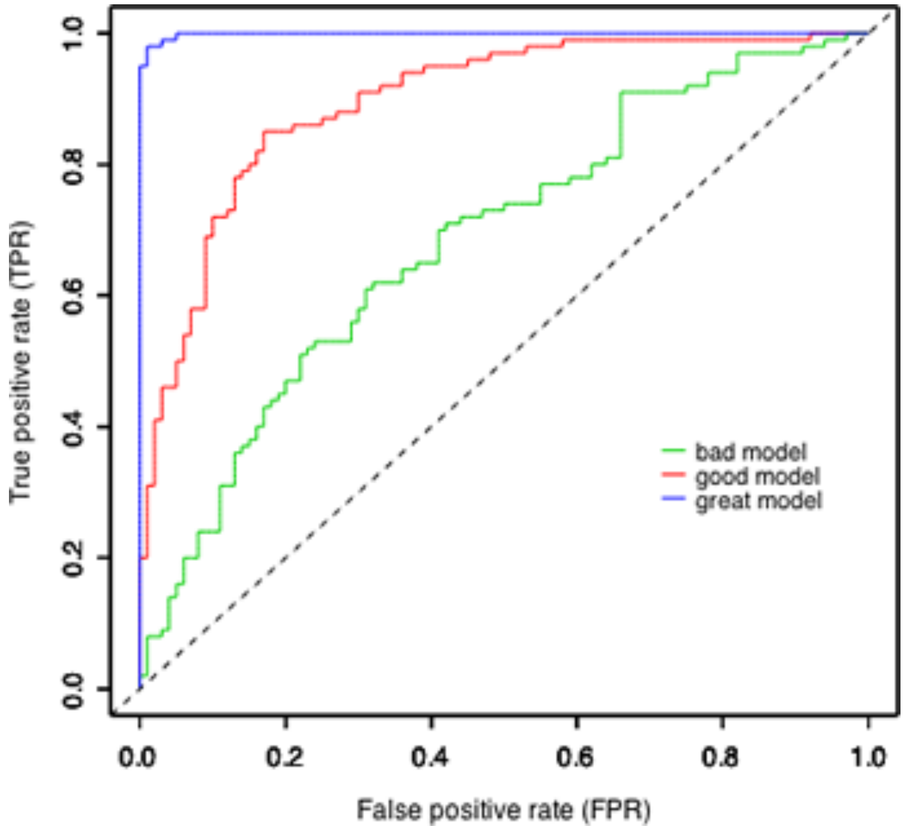
TPR = 75%
FPR = 17%

	Predicted +	Predicted -	Total
Sick +	39	1	40
Sick -	21	39	60
Total	60	40	100

TPR = 98%
FPR = 35%

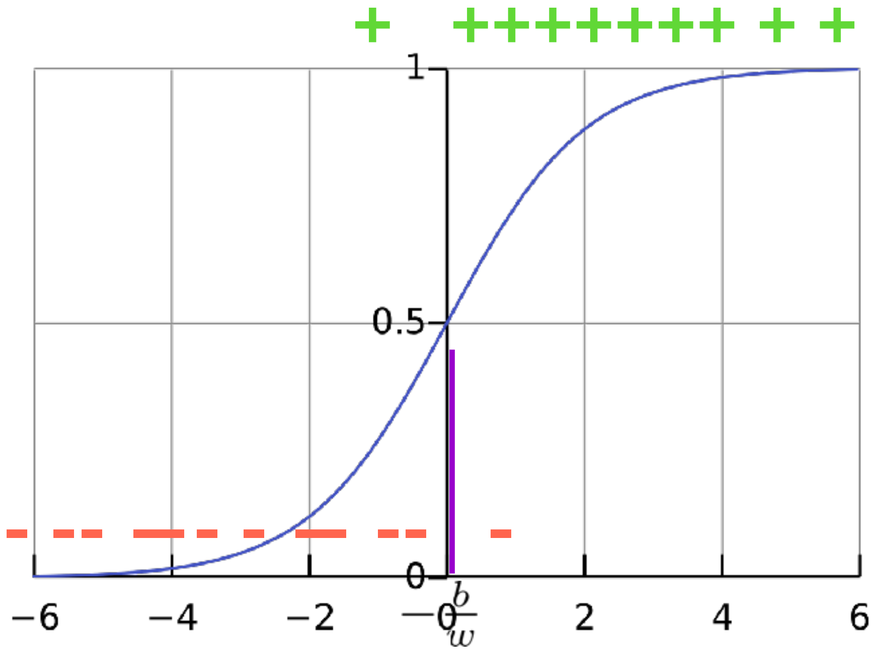
Receiver Operating Characteristic

ROC-AUC



The value can range from 0 to 1. However AUC score of a random classifier for balanced data is 0.5

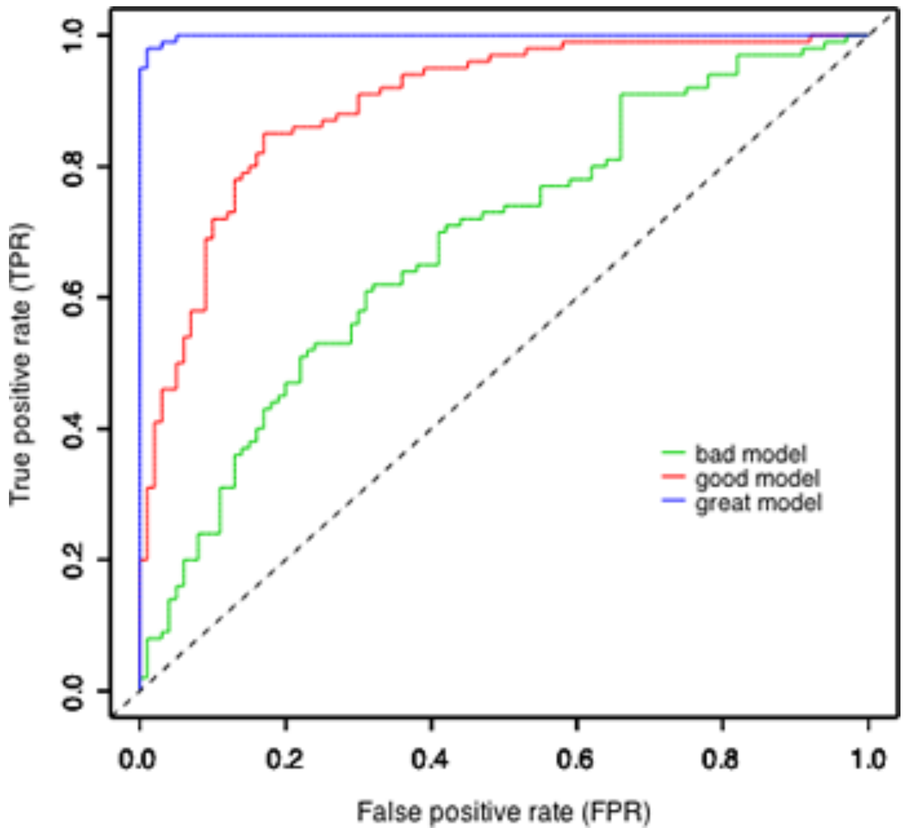
Courtesy of Alvira Swalin



We have: $f(x; w, b) = \begin{cases} +1 & if\ w \times x + b \geq 0 \\ -1 & otherwise \end{cases}$.

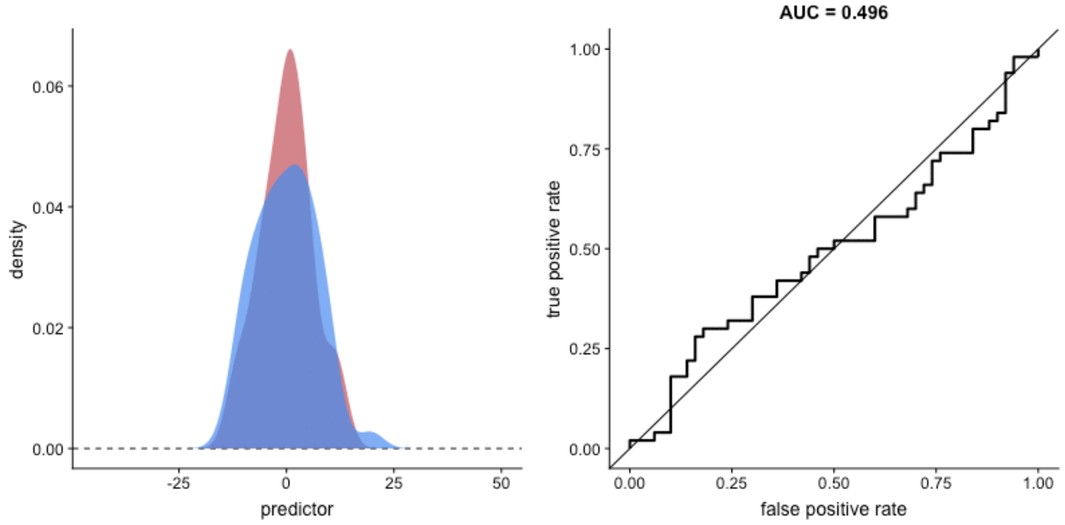
Receiver Operating Characteristic

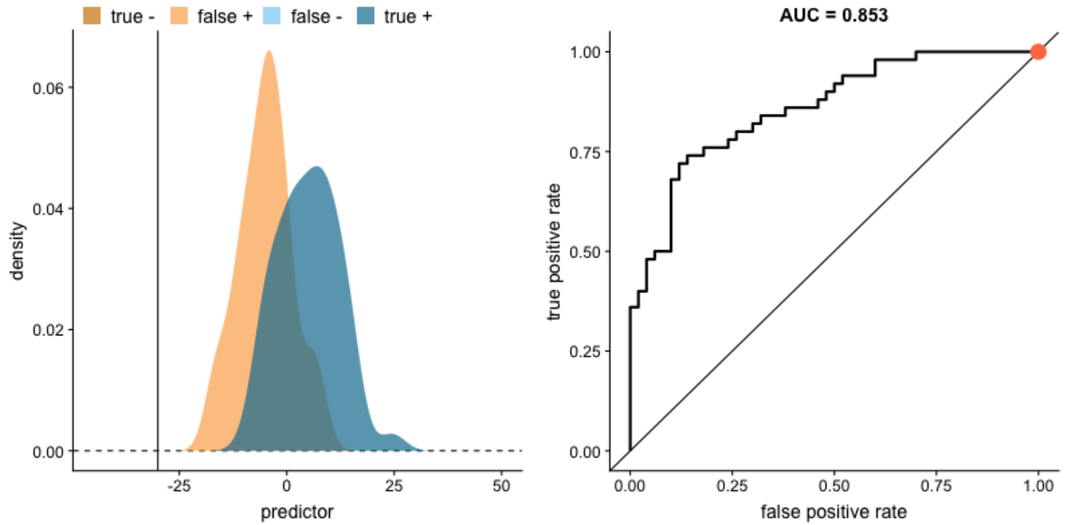
ROC-AUC



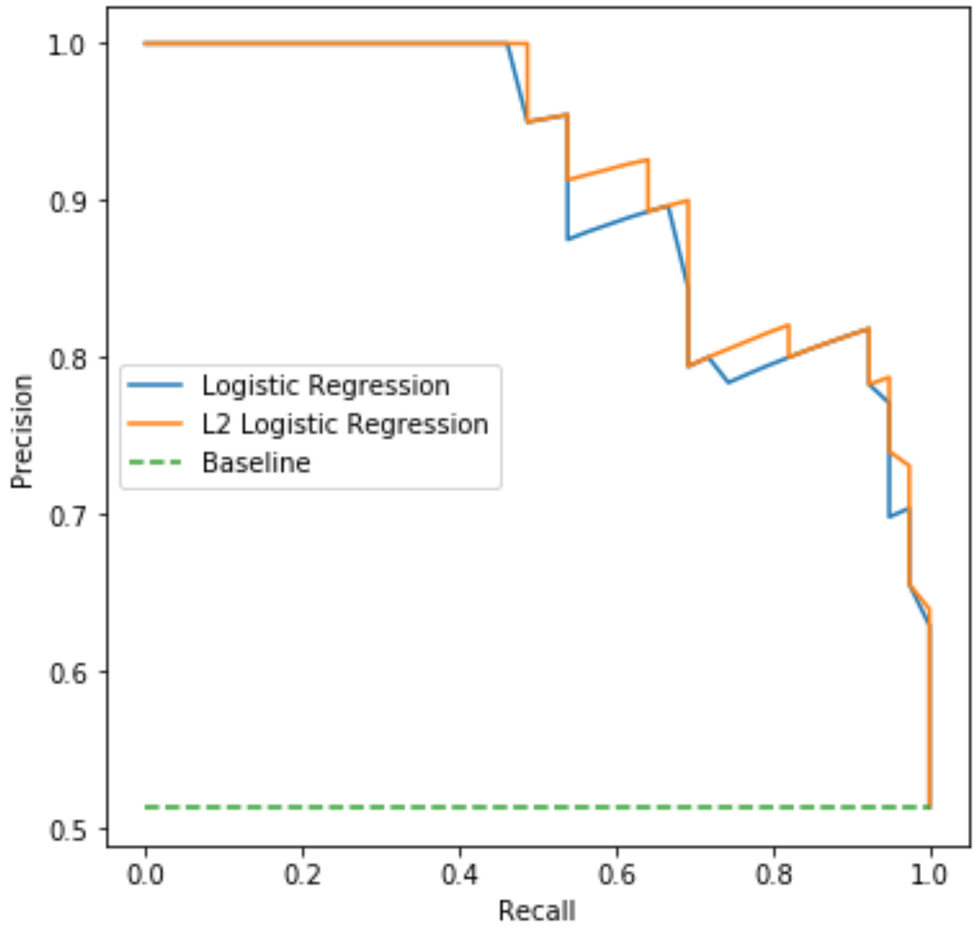
The value can range from 0 to 1. However AUC score of a random classifier for balanced data is 0.5

Courtesy of Alvira Swalin

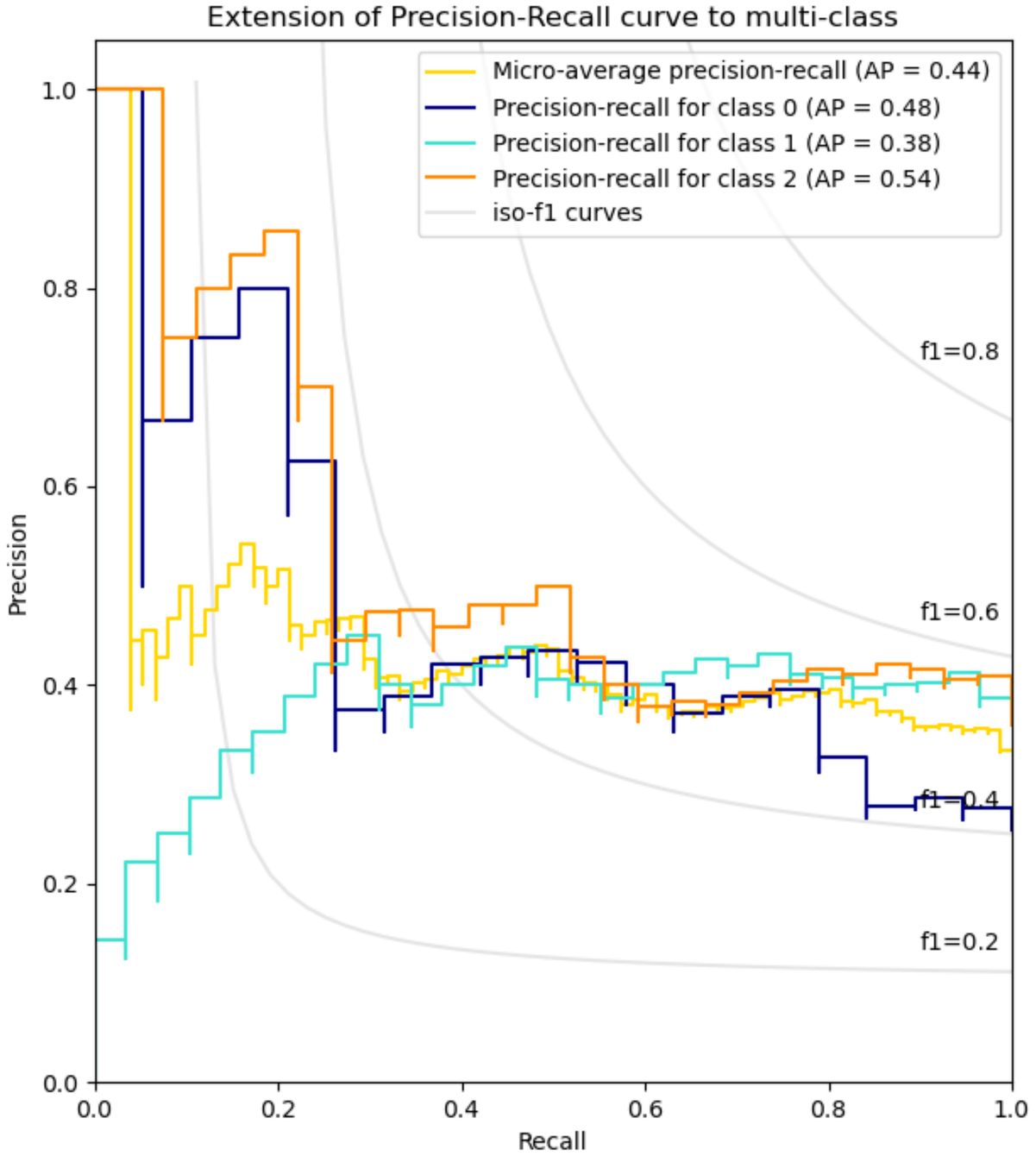




Precision - Recall curves



		Predicted condition		
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	


Extension of Precision-Recall curve to multi-class
Micro-average precision-recall (AP = 0.44)
Precision-recall for class 0 (AP = 0.48)
Precision-recall for class 1 (AP = 0.38)
Precision-recall for class 2 (AP = 0.54)
iso-f1 curves
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Whadda bout regression??

A gentle intro to R-squared

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

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

In the best case, the modeled values exactly match the observed values, which results in $SS_{\text{res}} = 0$ and $R^2 = 1$. A baseline model, which always predicts $\bar{y}$, will have $R^2 = 0$.

A problem… adding more features/model complexity always increases R-squared…
so there's a penalized version called Adjusted R-squared, also AIC/BIC, etc
OTHER common regression scores: MSE / RMSE, MAE