Logistic Regression – Assessment of Classification models

Dr. Mohamed Elshenawy

mmelshenawy@gmail.com

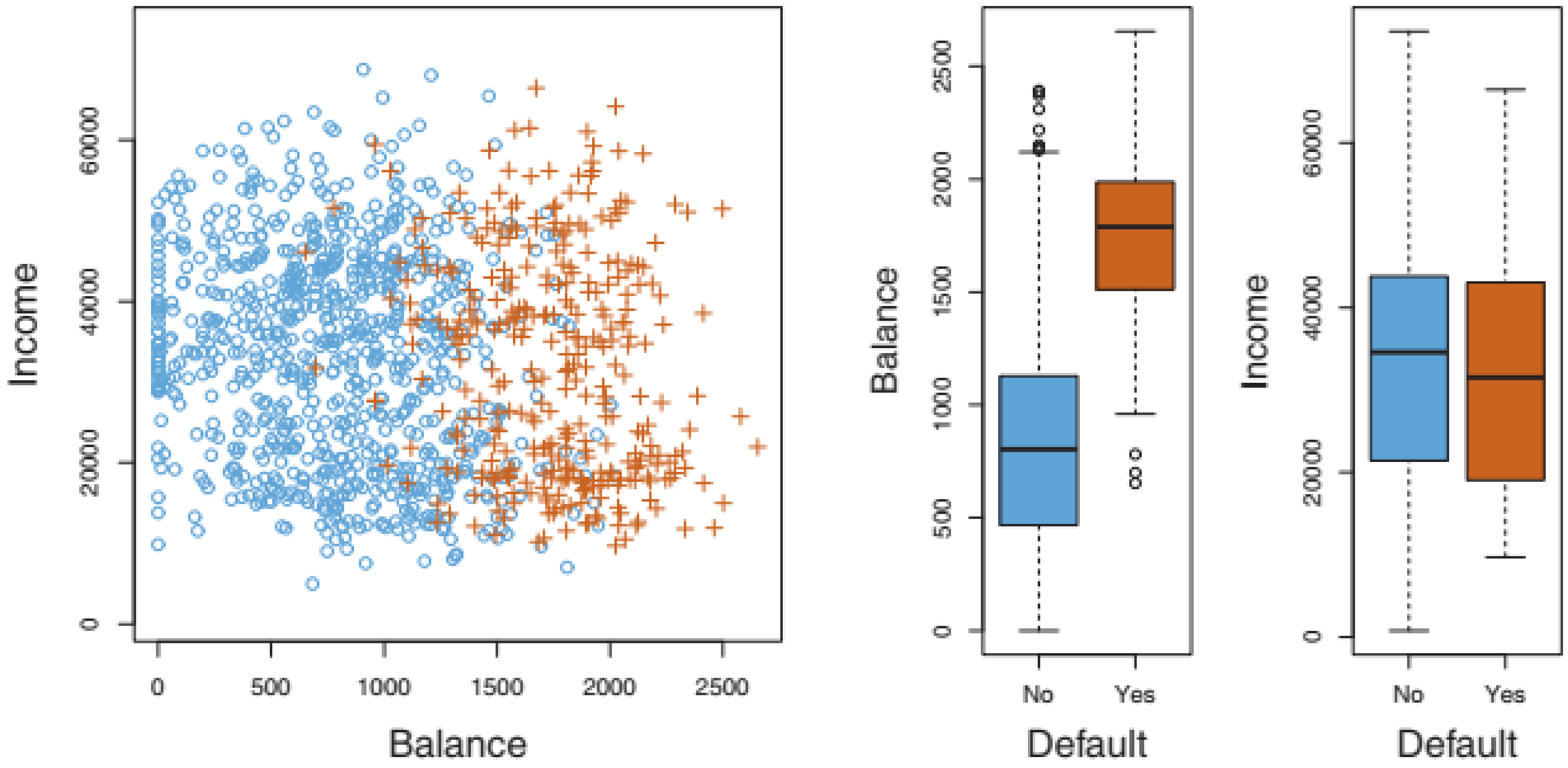
Previous sessions

- What is Learning?
- Regression - OLS
- Overfitting and underfitting
- Assessment of Regression Models
- Parametric and non-parametric models
- KNN

This session

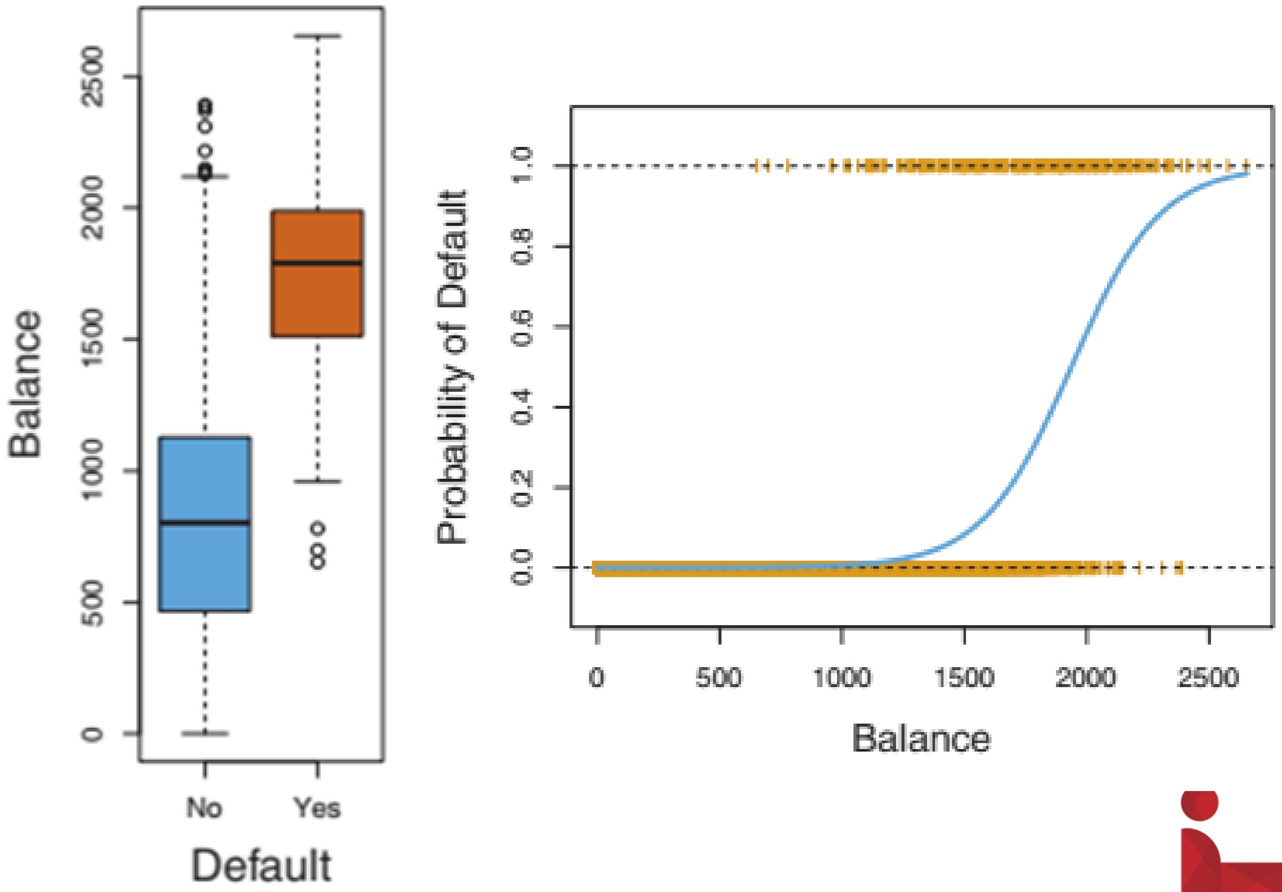
- Review of Assignment 1 (Example solution)
- Logistic Regression (cont.)
  - Maximum Likelihood Estimation
  - Multi-class classification
- Assessment of Classification Models

Previous session



In the credit card payments example

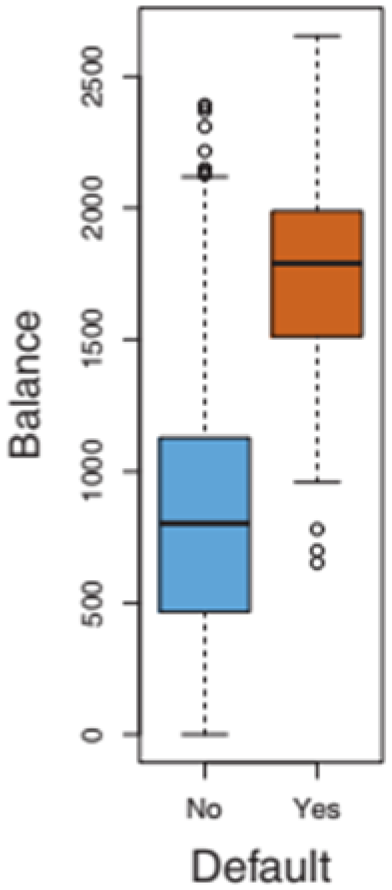
- $\Pr(default = Yes|blanace) = \sigma(\theta_0 + \theta_1 * balance)$
- One might predict default = Yes for any individual for whom $\Pr(default = Yes|blanace) > 0.5$.
- Alternatively, a more conservative approach in predicting whether an individual will default (fail to repay) on his or her credit card payment $\Pr(default = Yes|blanace) > 0.1$

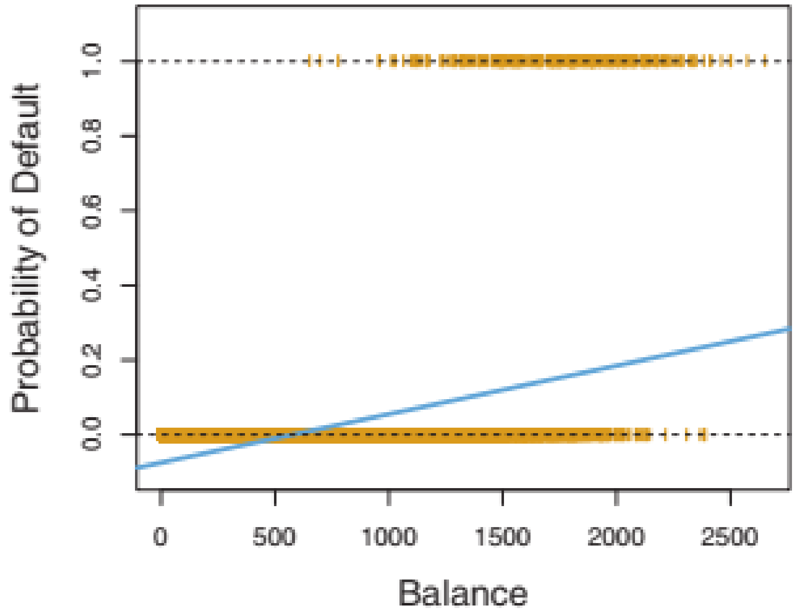


An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

Assume you use a linear model (without a logistic function)

- $\Pr(default = Yes|blanace) = \theta_0 + \theta_1 * blance$
- For balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1 (not sensible).

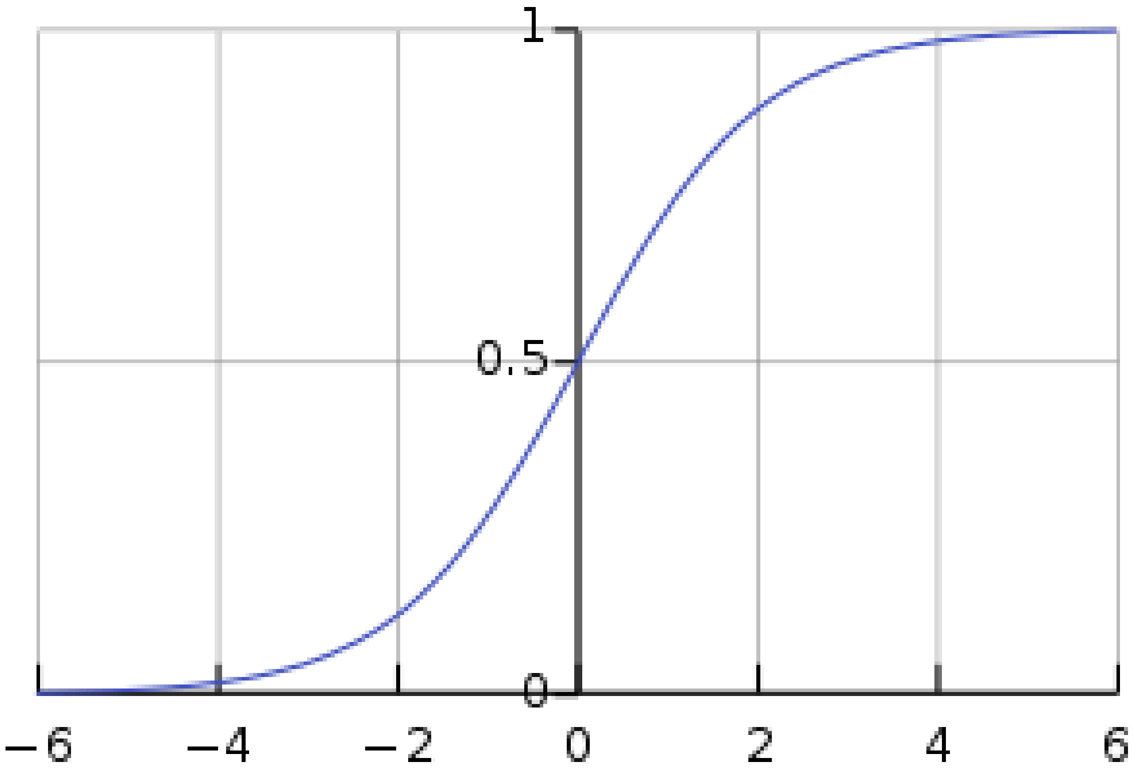




An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

Logistic Function

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{e^z + 1}$$



How can we estimate the parameters?

- $\Pr(Y = 1|X) = \sigma(\theta_0 + \theta_1 * X_1 + \theta_2 * X_2 + \cdots)$

How to estimate parameters: Maximum Likelihood Estimation

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{N} P(y^{(i)} = t^{(i)} | x^{(i)}; \theta)$$

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} log\ P(y^{(i)} = t^{(i)} | x^{(i)}; \theta)$$

- The discussion of MLE is out of the scope of this session.

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

How to estimate parameters: Maximum Likelihood Estimation

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{N} P(y^{(i)} = t^{(i)} | x^{(i)}; \theta)$$

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} log\ P(y^{(i)} = t^{(i)} | x^{(i)}; \theta)$$

$$gradient = \sum_{i=1}^{N} x_j^{(i)} \left(t^{(i)} - P(y^{(i)} = 1 | x^{(i)}; \theta)\right)$$

- The discussion of MLE is out of the scope of this session.

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

Multiclass Logistic Regression

- How about if we have more than two classes?
- For example, a person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions: stroke, drug overdose, epileptic seizure.
- In this setting, we wish to model:
  - $\Pr(Y = stroke|X)$
  - $\Pr(Y = drug\ overdose|X)$
  - $\Pr(Y = epileptic\ seizure|X)$
- The sum of the three probabilities is 1
- The two-class logistic regression models have multiple-class extensions (not popular).
- Later in ML2, we will discuss a more popular method (discriminant analysis methods)

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

How can we separate multiple classes?

Create 3 training sets, 3 models

Predict the class using the three models, compare

Winner is the selected class

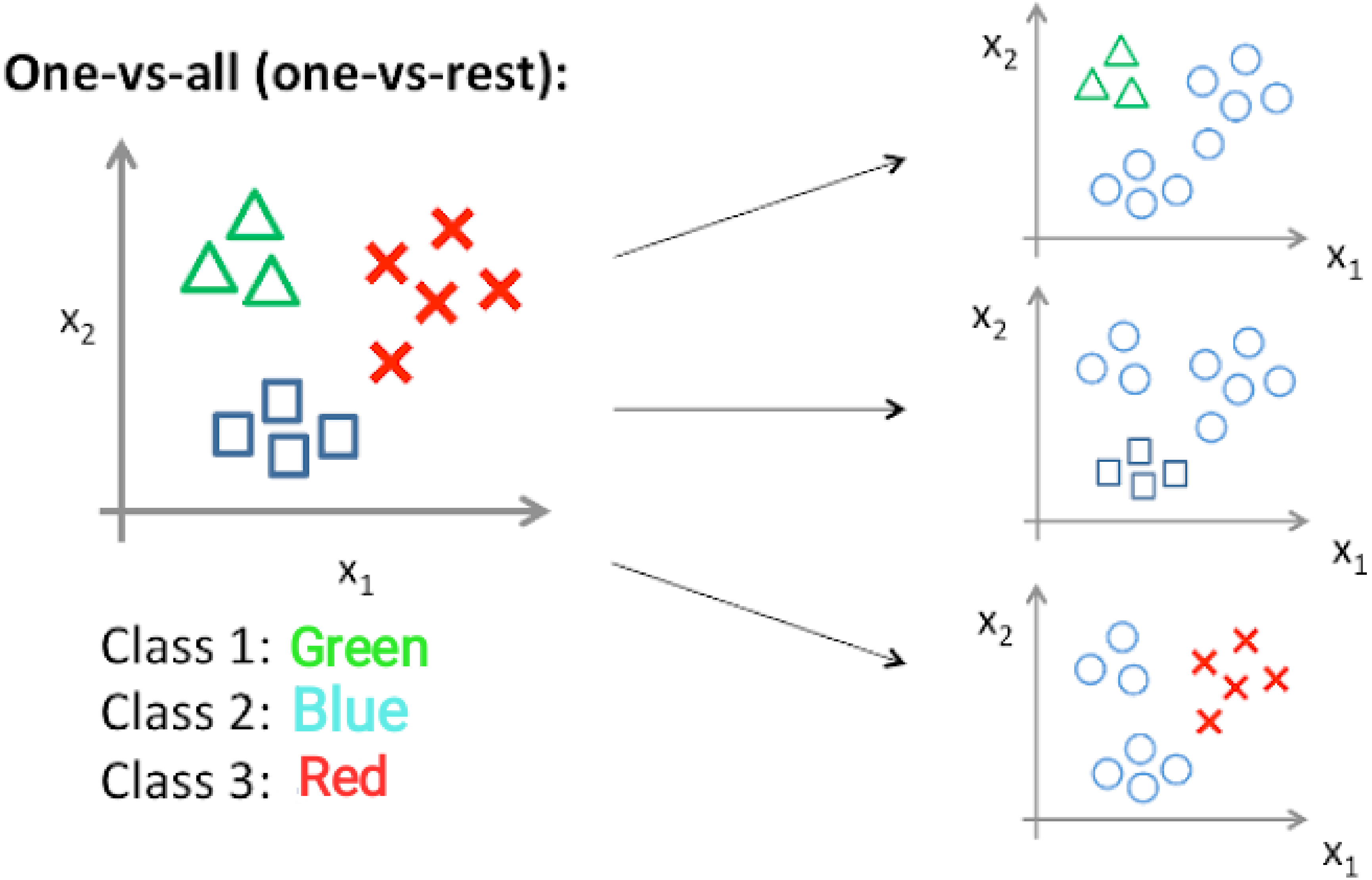
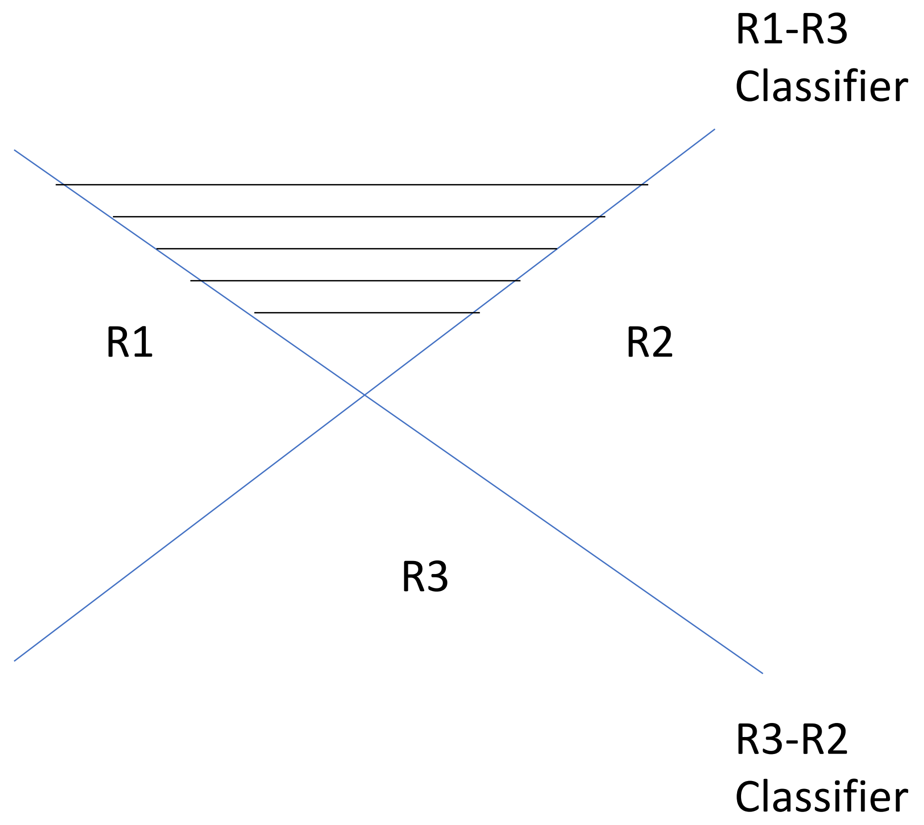


Photo via cc.gatech.edu

Use K-1 classifiers??

- Given K classes where K>2, how many lines do we need to separate these classes?
- An idea: we use K-1 classifiers, each separates two classes?
- How can classify the points in the shaded area?

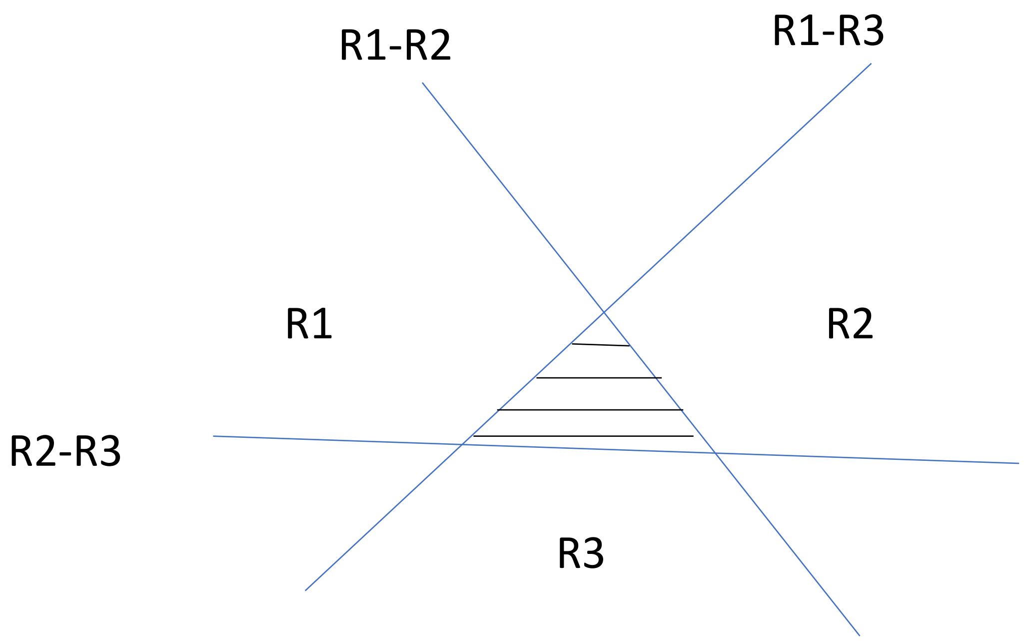


Use a classifier for each possible pair of classes (One-vs-One)

- Another idea: Use a classifier for each possible pair of classes. Fits (K-1) classifier per class.
- Number of lines = number of possible pairs

$$Number\ of\ lines = {}^{k}C_2 = \frac{k!}{(k-2)!\,2!} = \frac{k(k-1)}{2}$$

- How do you classify the points in the shaded area?



An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

K-Class Discriminant

- For K classes, use K functions

$$z_k(x) = \theta_k^T x + \theta_{k0}$$

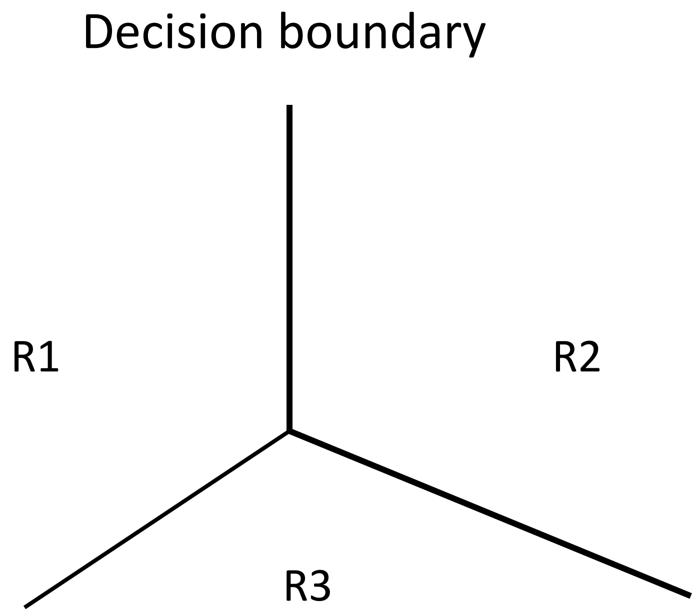
- Assign a point x to class k if

$$z_k(x) > z_j(x) \; for \; all \; j \neq k$$

- The decision boundary between two classes k, j is given by

$$z_k(x) = z_j(x)$$

that is $(\theta_k - \theta_j)^T x + (\theta_{k0} - \theta_{j0}) = 0$



Representation of the output: One Hot Encoding

	stroke	*drug overdose*	*epileptic seizure*
stroke	1	0	0
drug overdose	0	1	0
epileptic seizure	0	0	1

Softmax function

- We use softmax: a normalized exponential function

$$p(Y = k|x) = softmax(z_k) = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

Where

$$z_k = \theta_k^T x + \theta_{k0}$$

Normalized exponential function guarantee that the sum of probabilities is equal to 1 (not the case if a sigmoid function is used)

Using Maximum Likelihood

- The likelihood

$$\prod_{n=1}^{N}\prod_{k=1}^{K} p(Y = k|x^{(n)})^{t_k^{(n)}}$$

Using summation

$$L(\theta, \dots, \theta_k) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^{(n)} \log p(Y = k|x^{(n)})$$

Use gradient descent

$$\frac{\partial E}{\partial w_{k,i}} = \sum_{n=1}^{N}\left(\left(t_k^n - p(Y = k|x^{(n)})\right)\right) * x_i^{(n)}$$

Assessment of Classification Models

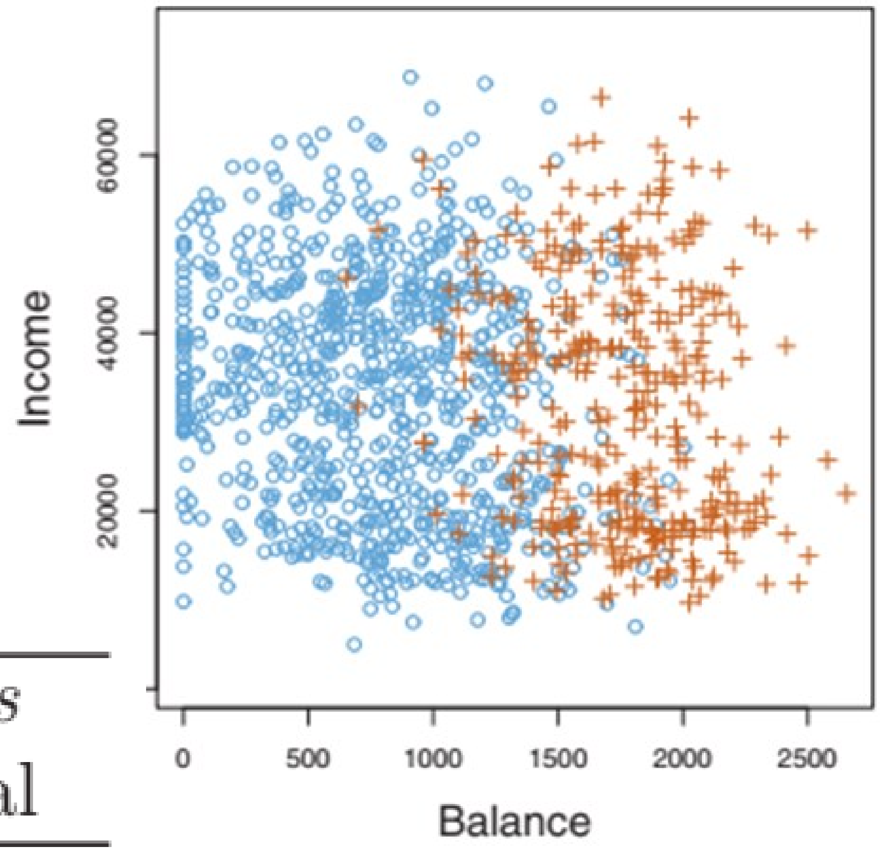
- We talked about accuracy: percentage of examples that are classified correctly by your model.
- Sometimes, accuracy is not sufficient to describe the model.
- Consider a diagnostic tool that uses a machine learning model to tell that person X has a disease Y.
- In this scenario, we are interested in two types of errors:
  - If X has the disease, what is the probability that the device say that X does not have it.
  - If X does not have this disease, what is the probability that the device will say that he has it.
- Consider an accident detection alarm
  - If the alarm is on, what is the probability that there’s an accident
  - If there’s an accident, what is the probability that the device will not detect it.

Confusion Matrix

Output of the classifier

	P	N
P	True positive (TP)	False Negative (FN)
N	False positive (FP)	True Negative (TN)

Example



		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Recall, Precision, F1-Score

- Recall (other names sensitivity and probability of detection) :

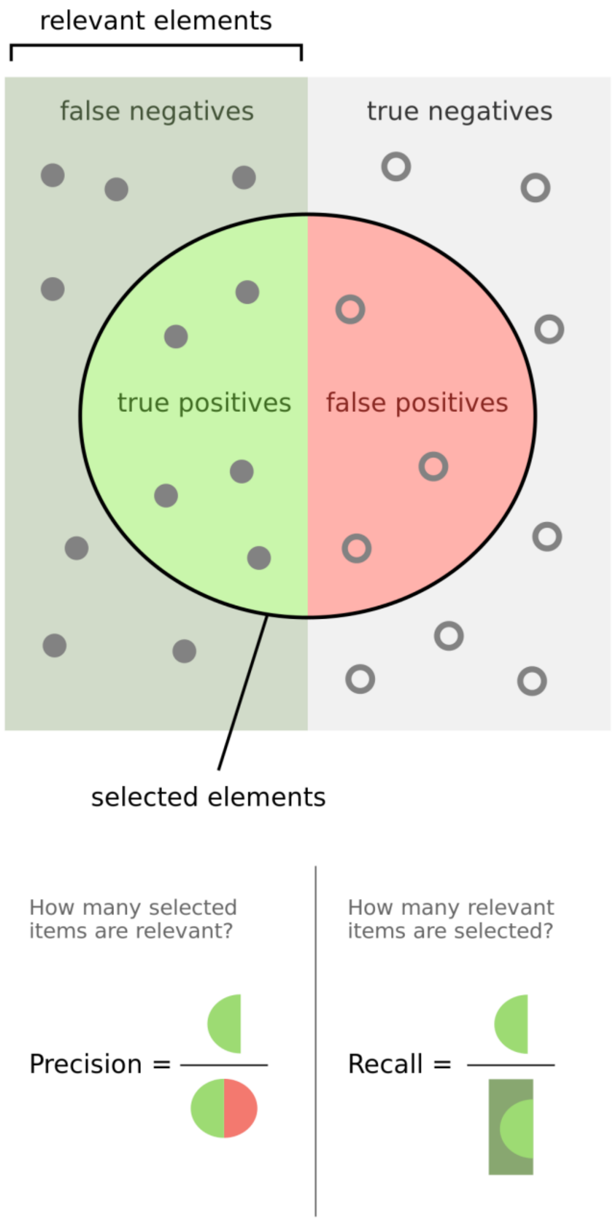
$$Recall\ (R) = \frac{TP}{TP + FN} = \frac{Retrieved}{All\ \text{groundtruth}\ positives}$$

- Precision

$$Precision(P) = \frac{TP}{TP + FP} = \frac{Retrieved}{All\ Positive\ classifications}$$

- F1 score: harmonic mean of precision and recall

$$F1 = 2\frac{P * R}{P + R}$$



Sensitivity and Specificity

- Common metrics in biology
- Sensitivity (same as recall)

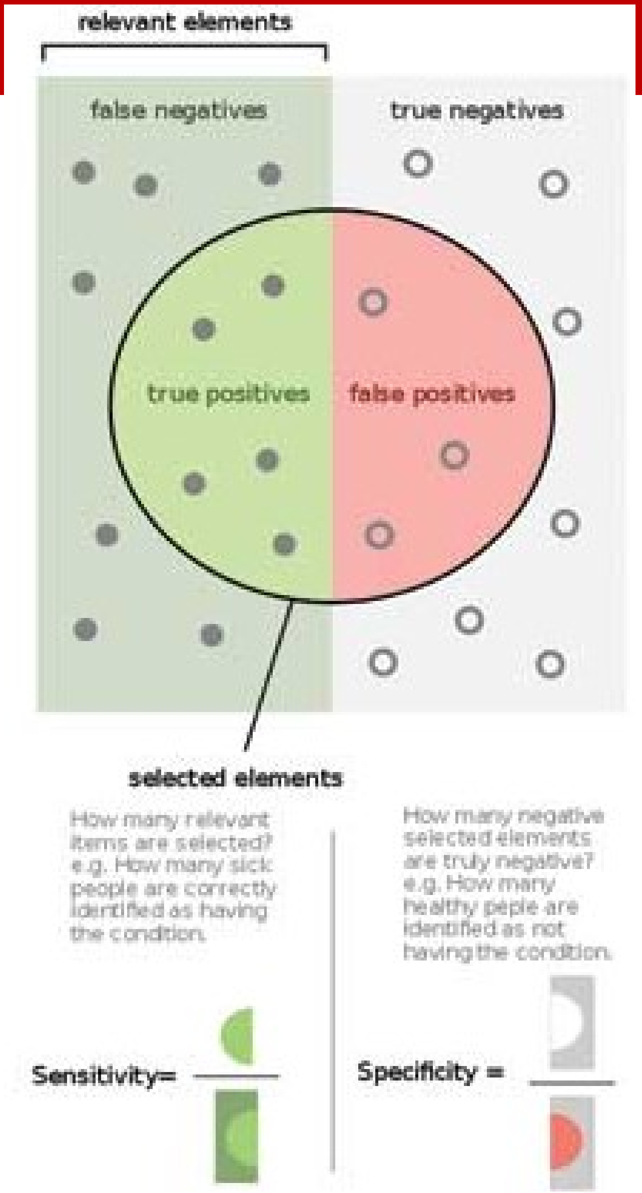
$$Sensitivity = \frac{TP}{TP + FN}$$

$$= \frac{Number\ of\ true\ positives}{Total\ number\ of\ cases\ with\ illness}$$

- Specificity

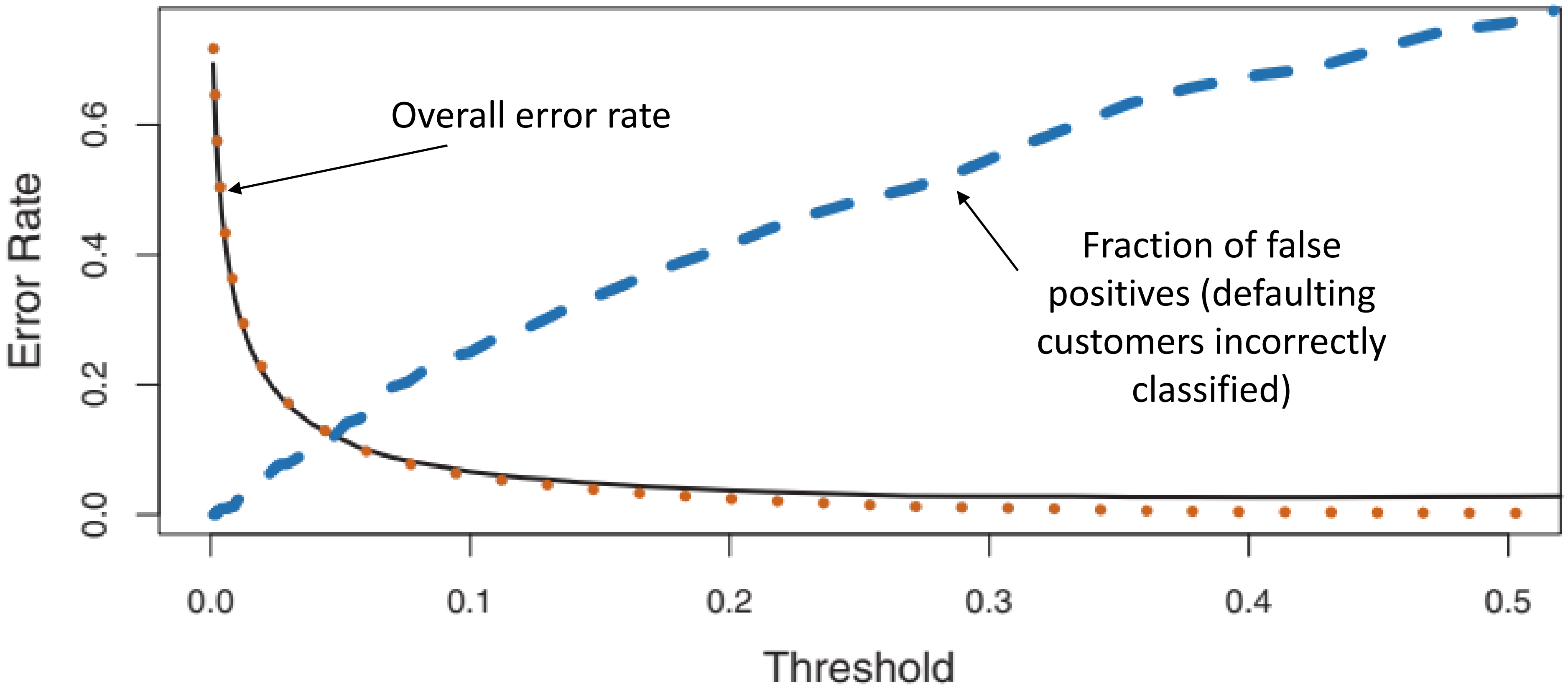
$$Specifcity = \frac{TN}{FP + TN}$$

$$= \frac{Number\ of\ true\ negatives}{Total\ number\ of\ cases\ without\ the\ illness}$$



Choosing different thresholds

- Can we use thresholds to reduce the number of false positive?
- If P(Y=1|X=x) greater than 0.5, does it improve the performance



ROC Curve

- The ROC curve is a popular graphic for simultaneously displaying false positives and true positives for all possible thresholds.
- he name “ROC” comes from communications theory. It is an acronym for receiver operating characteristics.
- The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (AUC).
- An ideal ROC curve will hug the top left corner, so the larger area under the AUC the better the classifier.