MAIN CONCEPTS CLASSIFICATION

NOTATION

- Input variables: $X = (X_1, X_2, \ldots, X_p)$ - independent variables, predictors, features
- Output variable: Y - response, dependent variables – categorical data

$$Y = j, \qquad j = 1,2, \ldots, K$$

- Classifier is a function (mapping) C

$$C: X \to Y, \qquad C: R^p \to \{1,2, \ldots, K\}$$

- We assume a law

$$Y = C(X) + e, \qquad E[e] = 0$$

- We should predict Y using

$$\hat{Y} = \hat{C}(X)$$

where $\hat{C}$ is estimate of C and $\hat{Y}$ is the prediction of Y

ERROR RATE OF CLASSIFICATION

- Assume classification problem with **training data**

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

where Y_k are some classes

- **Training accuracy**

$$Accuracy = \frac{1}{n}\sum_{k=1}^{n} I(Y_k = \hat{Y}_k)$$

- **Training error rate**

$$Error\ rate = 1 - Accuracy = \frac{1}{n}\sum_{k=1}^{n} I(Y_k \neq \hat{Y}_k)$$

- Also known as mean misclassification error (MME)

TEST ERROR RATE OF CLASSIFICATION

- We are most interested in the error rates that result from applying our classifier to test observations that were not used in training
- This is known as **test error rate**
- A **good classifier** is one for which the test error rate is the smallest

PROBABILITY SETTING - BAYES CLASSIFIER

- Probability of misclassification

$$R(C) = P\{\hat{C}(X) \neq Y\}$$

- How to minimize the test error?
- Calculate the following conditional probabilities

$$\Pr(Y = j|X = x_0)\,, j = 1,2, \dots, K$$

and

$$C^{Bayes}(X) = \underset{j\in\{1,2,\dots,K\}}{argmax} \Pr(Y = j|X = x_0)$$

- Bayes classifier **assigns each observation to the most likely class**
- This very simple classifier is called the ***Bayes classifier***
- The Bayes classifier produces the lowest possible test error rate, called the **Bayes error rate**

BAYES CLASSIFIER: CONTINUED

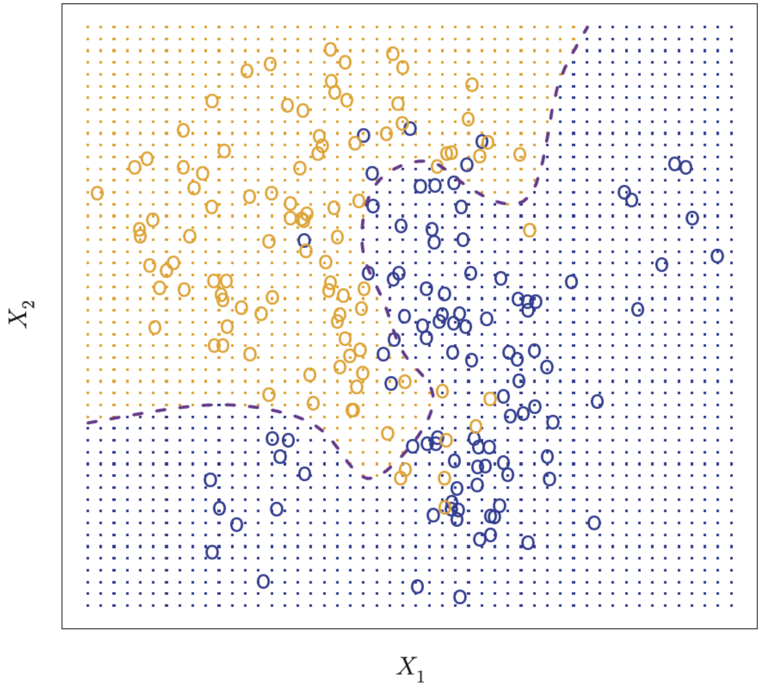
- The Bayes classifier is a useful benchmark in classification
- The following non-negative quantity (excess risk)

$$R(C) - R^{Bayes}(C)$$

is important for assessing the performance of different classification techniques

- A classifier is said to be consistent if the excess risk converges to zero as the size of the training data set tends to infinity

EXAMPLE



- Binary classification problem – orange and blue circles
- For this simulated data, we can calculate
$$P_j = \Pr(Y = j|X = (x_1, x_2)),$$
$$j = orange, blue$$
- The orange shaded region consists of the points for which
$$P_{orange} > 0.5\ (P_{blue} \leq 0.5)$$
- The blue shaded region consists of the points for which
$$P_{blue} > 0.5\ (P_{orange} \leq 0.5)$$
- The dashed line consists of points
$$P_{orange} = P_{blue} = 0.5$$
- This is called the **Bayes decision boundary**
- The Bayes classifier's prediction is determined by the Bayes decision boundary; an observation that falls on the orange side of the boundary will be assigned to the orange class …

CONFUSION MATRIX FOR A BINARY CLASSIFICATION

$n = N + P$

		Actual Classes	
		Positive Class P	Negative Class N
Predicted Classes	Positive Class	True Positives TP	False Positives FP
	Negative Class	False Negatives FN	True Negatives TN

CLASSIFICATION MEASURES

$$Accuracy = \frac{TP + TN}{n}$$

$$Precision = Positive\ Predictive\ Value(PPV) = \frac{TP}{TP + FP}$$

$$True\ Positive\ Rate\ (TPR) = Sensitivity = Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$False\ Positive\ Rate(FPR) = Fall\ Out = \frac{FP}{FP + TN} = \frac{FP}{N}$$

PRECISION VS RECALL

- High precision means that an algorithm returned substantially more relevant results than irrelevant (usefulness), while high recall means that an algorithm returned most of the relevant results (completeness).
- **Example:** Search engine returns 30 pages where only 20 pages are relevant ($TP = 20$) while failing to return 40 additional relevant pages ($FN = 40$). Its $precision = 20/30$ while its $recall = 20/60$.
- So, in this case, precision is "how useful the search results are", and recall is "how complete the results are".

CLASSIFICATION MEASURES

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{n}$$

$$Balanced\ Accuracy = \frac{\left(\frac{TP}{P} + \frac{TN}{N}\right)}{2} = \frac{Sensitivity + Specificity}{2}$$

$$Cohen's\ kappa = \frac{Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

$$Expected\ Accuracy \sim Random\ Classifier\ Accuracy$$

CLASSIFICATION MEASURES

$$Prevalence \sim percentage\ of\ each\ class$$

$$Prevalence = \frac{P}{n}$$

$$Detection\ Prevalence = \frac{TP + FP}{n}$$

$$Detection\ Rate = \frac{TP}{n}$$

CONFUSION MATRIX FOR A MULTICLASS CLASSIFICATION

$$n = A + B + C$$

Actual Classes / Predicted Classes

Predicted \ Actual	A	B	C
A	TP_A	BA	CA
B	AB	TP_B	CB
C	AC	BC	TP_C

$$FPR_A = \frac{AB + AC}{B + C}$$

$$FPR_B = \frac{BA + BC}{A + C}$$

$$FPR_C = = \frac{CA + CB}{A + B}$$

$$Accuracy = \frac{TP_A + TP_B + TP_C}{n}$$

$$TPR_A = \frac{TP_A}{A}$$

$$TPR_B = \frac{TP_B}{B}$$

$$TPR_C = \frac{TP_C}{C}$$

$$Precision_A = \frac{TP_A}{TP_A + FP_A} = \frac{TP_A}{TP_A + BA + CA}$$

$$Precision_A = \frac{TP_B}{TP_B + FP_B} = \frac{TP_B}{TP_B + AB + CB}$$

$$Precision_C = \frac{TP_C}{TP_C + FP_C} = \frac{TP_C}{TP_A + AC + BC}$$

BASELINE - MAJORITY CLASS CLASSIFIER

Actual Class

n		P	N
Predicted Class	P	0	0
	N	FN	TN

- Assume that the majority is the negative class
- x is the fraction of positives and $1 - x$ is the fraction of negatives:
  $$P = x\,n, \qquad N = (1 - x)n$$
- $FN = xn,$
- $TN = (1 - x)n$
- $TP = FP = 0$
- $Accuracy = 1 - x$
- $Precision = 0$
- $\boldsymbol{TPR = FPR = 0}$

BASELINE - RANDOM GUESS

		Actual Class	
	n	P	N
Predicted Class	P	$\frac{xn}{2}$	$\frac{(1-x)n}{2}$
	N	$\frac{xn}{2}$	$\frac{(1-x)n}{2}$

- Randomly assign half of the labels to positives and the other half to negatives
- x is the fraction of positives and $1-x$ is the fraction of negatives
- $TP = FN = \frac{xn}{2}$
- $TN = FP = \frac{(1-x)n}{2}$
- $Accuracy = \frac{1}{2}$
- $Precision = x$
- $\boldsymbol{TPR = FPR = \frac{1}{2}}$

BASELINE - WEIGHTED RANDOM GUESS

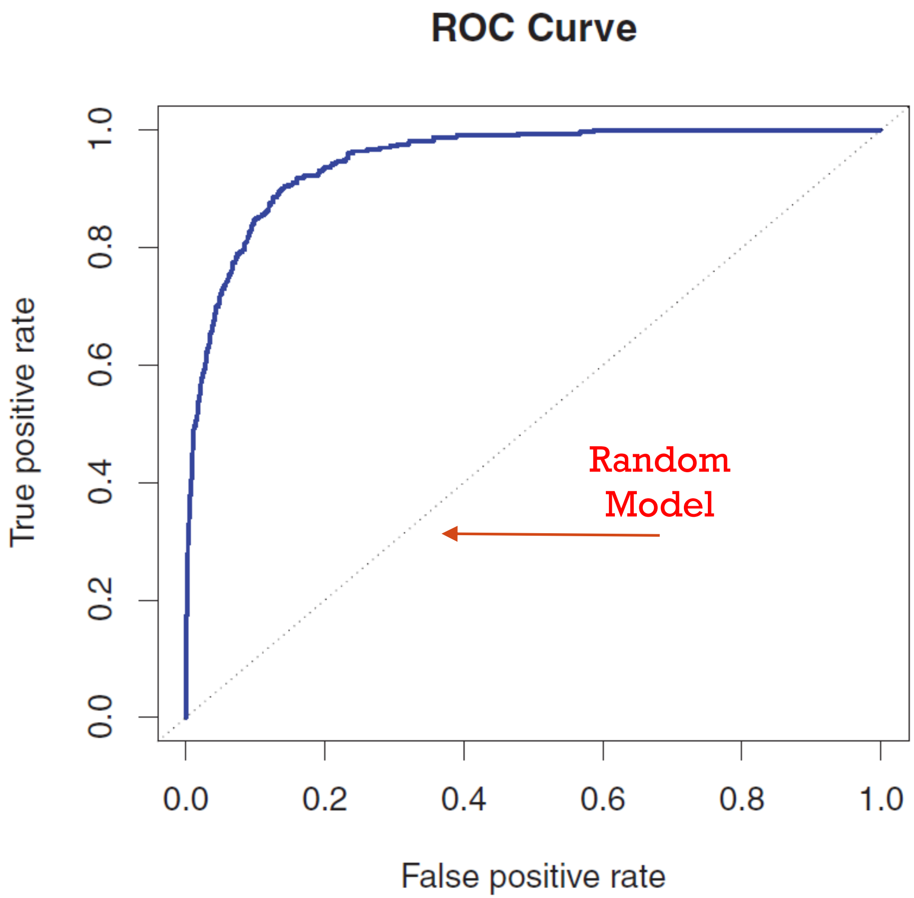
		actual	
	n	P	N
predicted	P	xP	$(1-x)P$
	N	xN	$(1-x)N$

- x is the fraction of positives
- randomly assign x portion to positives, and $1-x$ to negatives
- $TP = x^2 n$
- $FN = (1-x)xn$
- $TN = (1-x)^2 n$
- $FP = x(1-x)n$
- $Accuracy = x^2 + (1-x)^2$
- $Precision = x$
- $\boldsymbol{TPR = FPR = x}$

ROC CURVE

- ROC curve (*receiver operating characteristics) is a graph of TPR against FPR* for different thresholds
$$P(Y = j|X = x_0) = h$$
- ROC analysis provides tools to select possibly optimal models. Ideal classifier corresponds to the left-upper corner of the ROC curve with TPR = 1 and FPR = 0
- The overall performance of a classifier, summarized over all possible thresholds, is given by area under the curve (AUC)
- A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1

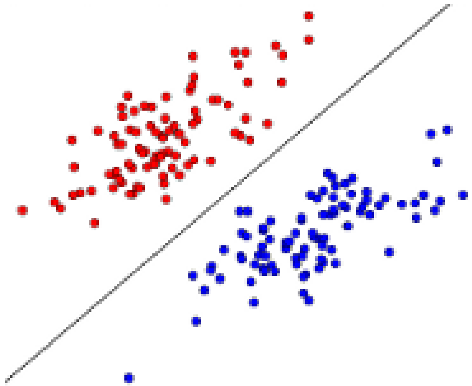
ROC CURVE

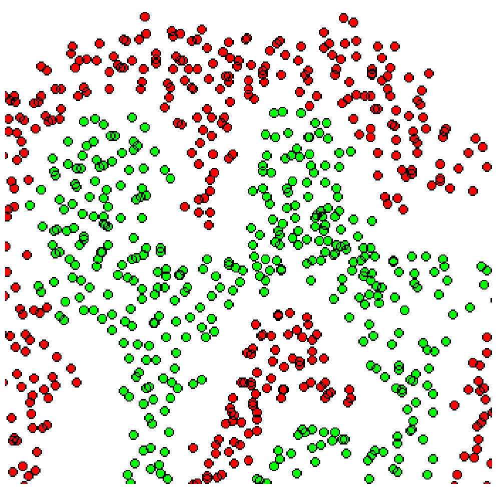

ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Random
Model

DECISION BOUNDARIES

- Any classification algorithm determines decision boundaries which separate identified classes. The classifier will classify all the points on one side of the decision boundary as belonging to one class and all those on the other side as belonging to the other class.

DECISION BOUNDARIES





- Two classes are linearly separable if it is possible perfectly classify them with a linear decision boundary.
- Linear classifiers can only draw linear decision boundaries.
- Non-linear classifiers have non-linear, and possibly discontinuous decision boundaries.

DECISION BOUNDARIES

Two sets are linearly separable if their convex hulls have no intersection.