DTU

02450 Introduction to Machine Learning and Data Mining

Week 9: AUC and ensemble methods

Bjørn Sand Jensen

1 April 2025

DTU Compute, Technical University of Denmark

Today

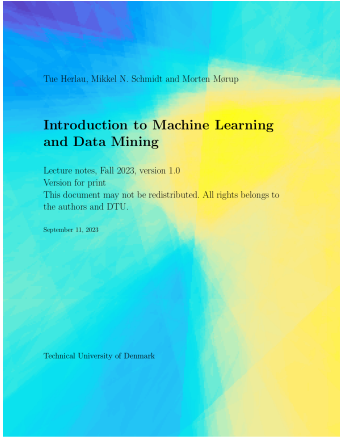
Feedback Groups of the day:

Theis Østergaard Hove, Zahra Soleimani, Sivalaxmanan Baleswaran Krishnapillai, Dan Howard Edelstein, William Kento Rasmussen, Jorge Ernesto Figueroa Vallecillo, Lucas Burmester, Katinka Clara Spangtoft, Anne Sofie Gadfelt, Liv Victoria Sjørup, Ziwei Li, Jakob Vincent Geishauser Hald, Benjámin Márk Telek, Oscar Wohlfahrt, Jacqueline Printz, Jakub Bouzan, Haseeb Shafi, Molly Jean Maud Turner, Karin van Keulen, Sarah Hecht Petersen, Martina Zini, Francisco Breia de Oliveira, Daniel Brinkmann, Magnus Lau Ovesen, Mattis Edelbo Kragh, Bastian Røder Clemmensen, Rishikesh Umamaheswaran, Xiaosa Liu, Anna Emilie Lunde Borre, Gregers Thomas Skat Rørdam, Oliver Jessen, Kal Adam Kalo, Sofia Theodora Thirslund, Joachim Rønsholt, Thea Rehm Rosholm, Zahraa Abdulrihman, Julie Thanh Thanh Nguyen, Hanne Daltveit, Eva Fasting Narvestad, Anna Victoria Spang Kollerup, Malthe Gilbert Jespersen, Alexandru Cecan

Reading/homework material:

Chapter 16, 17

P16.1, P16.2, P17.1



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. **AUC and ensemble methods**
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

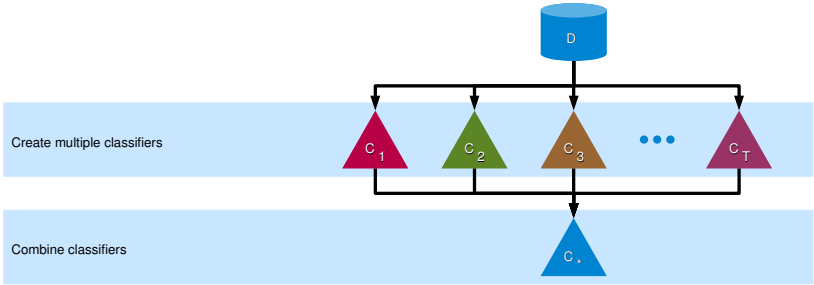
Recap

13. Recap and discussion of the exam
    6 May: C1-C21

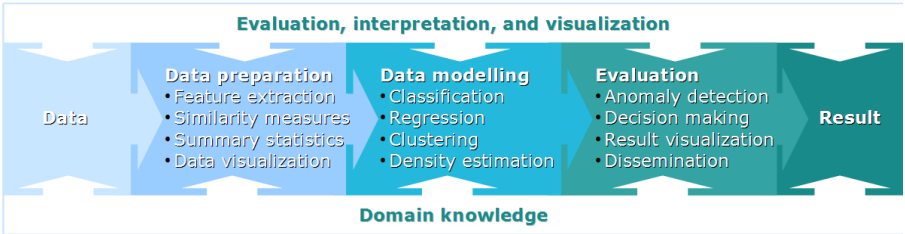
Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Ensemble methods

- Combine multiple (weak) classifiers into one (strong) classifier
- Each classifier trained using **different** variations of
  - Data set
  - Input attributes
  - Class label
  - Algorithm/model



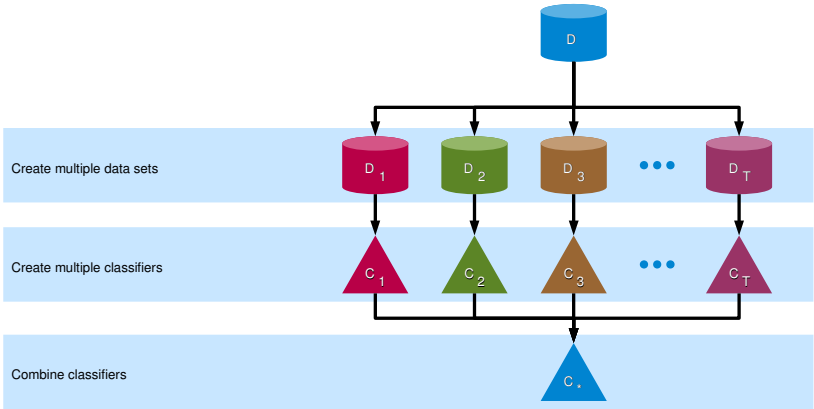
Learning Objectives



Learning Objectives

- Explain the principle behind boosting and bagging and apply it to improve classifiers
- Be able to address issues of class-imbalance and resampling
- Understand the definition of Precision, Recall, ROC, and AUC

Ensemble methods



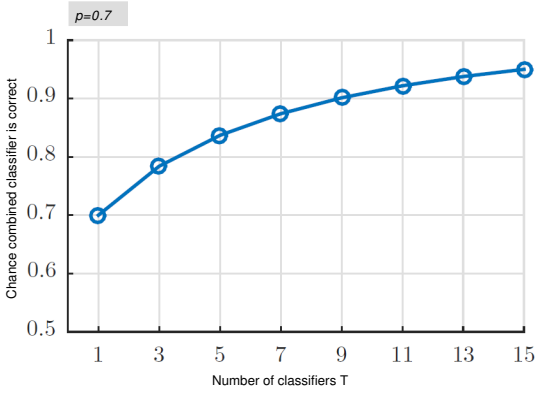
Why ensemble methods?

- Can improve classification algorithms in terms of
  - Better classification accuracy
  - Increased stability
  - Reduced variance
  - Less overfitting
- Consider T **independent classifiers** for binary classification, each with accuracy p.
- The probability a classifier which use majority voting is correct is then given by:

$$
\begin{aligned}
P(\text{Majority voting is correct}) &= \sum_{t=\lceil T/2 \rceil}^{T} P(\{t \text{ classifiers are correct}\}) \\
&= \sum_{t=\lceil T/2 \rceil}^{T} \binom{T}{t} p^t (1-p)^{T-t}
\end{aligned}
$$

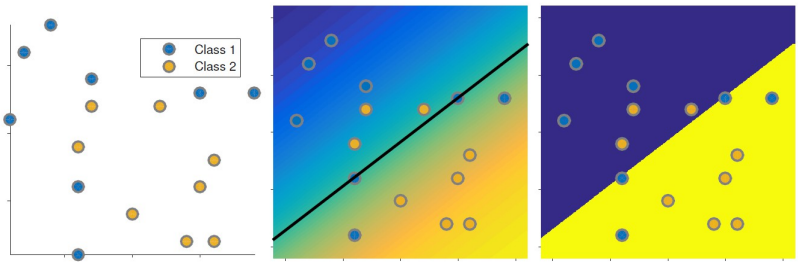
Why ensemble methods?

$$P(\text{Majority voting is correct}) = \sum_{t=\lceil T/2 \rceil}^{T} P(\{t \text{ classifiers are correct}\})$$
$$= \sum_{t=\lceil T/2 \rceil}^{T} \binom{T}{t} p^t (1-p)^{T-t}$$



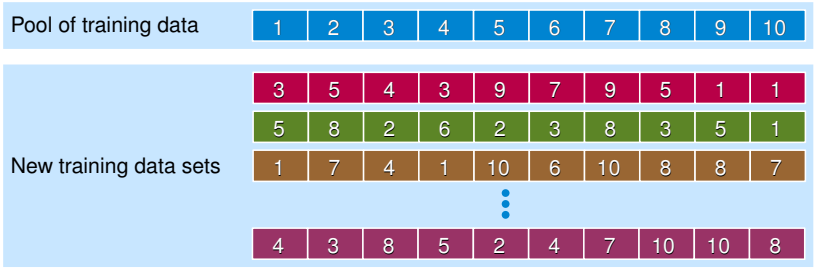
Recall: Logistic regression

- Classification using logistic regression

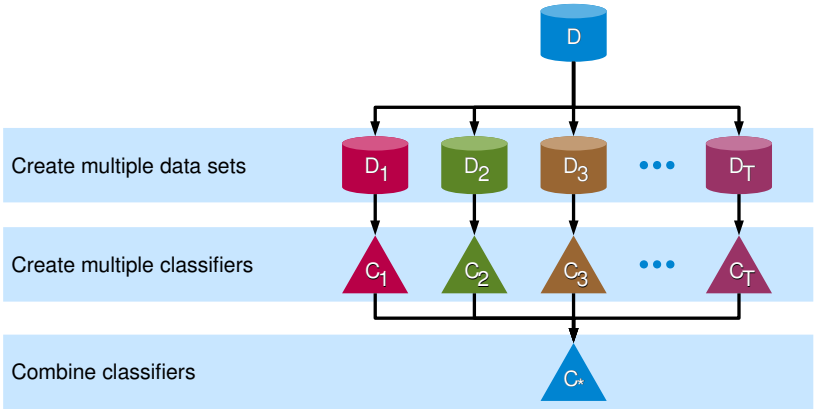


Bagging

- New training data sets drawn randomly from pool with replacement

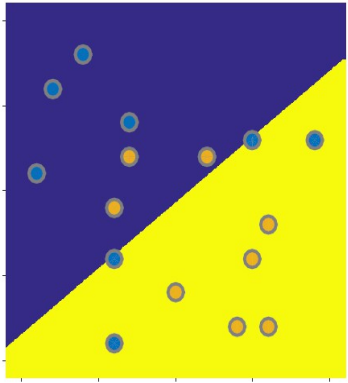


Bagging

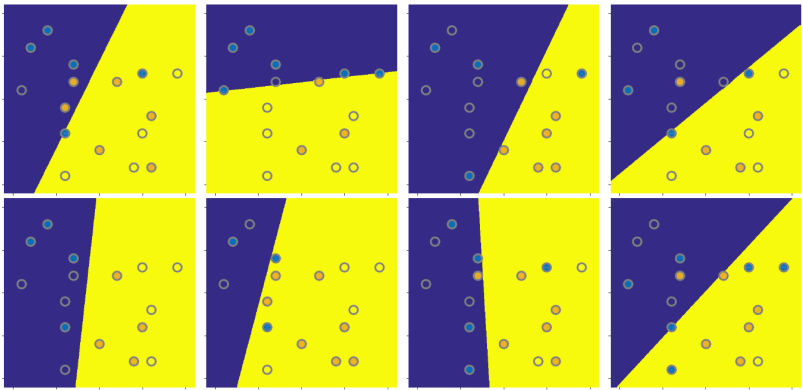

D
Create multiple data sets
D_1
D_2
D_3
D_T
Create multiple classifiers
C_1
C_2
C_3
C_T
Combine classifiers
C_*

Bagging

- **Single classifier**
  - Logistic regression
  - Two attributes (x_1,x_2)



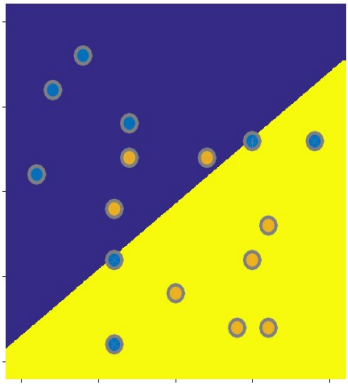
Bagging



Notice, hollow dots are observations not included in bagging round

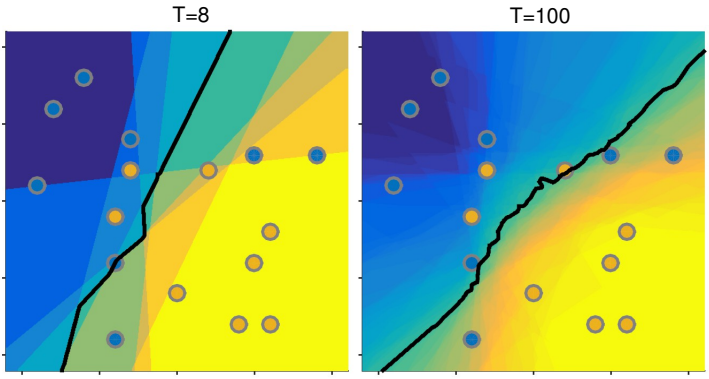
Bagging

- Recall: Single classifier

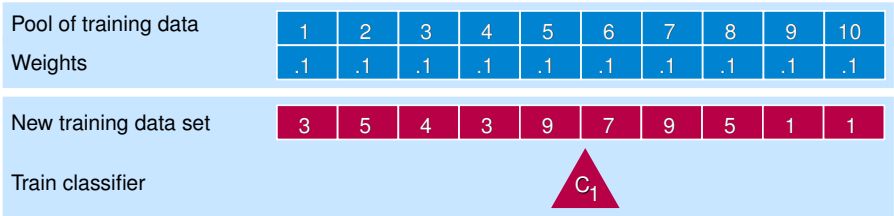


Bagging

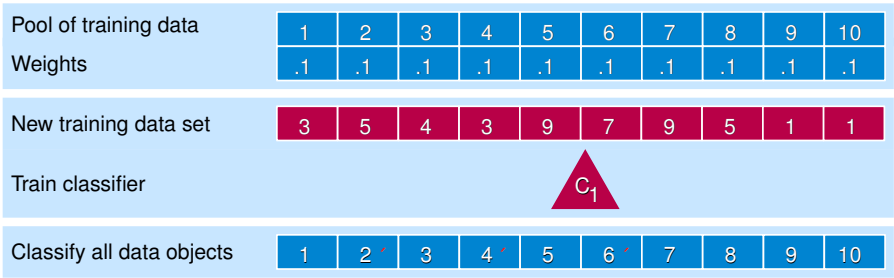
- Recall: Single classifier



Boosting


Pool of training data
1 2 3 4 5 6 7 8 9 10
Weights
.1 .1 .1 .1 .1 .1 .1 .1 .1 .1
New training data set
3 5 4 3 9 7 9 5 1 1
Train classifier
C_1

Boosting


Pool of training data
1 2 3 4 5 6 7 8 9 10
Weights
.1 .1 .1 .1 .1 .1 .1 .1 .1 .1
New training data set
3 5 4 3 9 7 9 5 1 1
Train classifier
C_1
Classify all data objects
1 2′ 3 4′ 5 6′ 7 8 9 10

Boosting

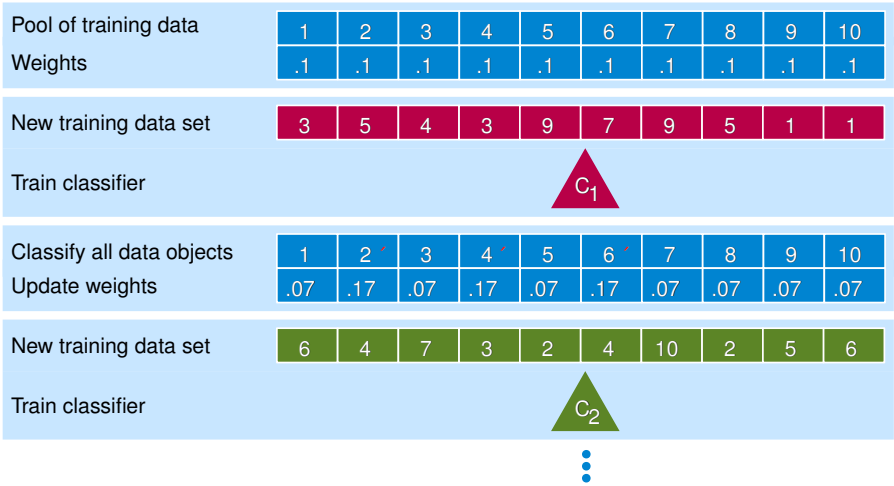
Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1

New training data set	3	5	4	3	9	7	9	5	1	1

Train classifier

Classify all data objects	1	2	3	4	5	6	7	8	9	10
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07

Boosting


Pool of training data
1 2 3 4 5 6 7 8 9 10
Weights
.1 .1 .1 .1 .1 .1 .1 .1 .1 .1
New training data set
3 5 4 3 9 7 9 5 1 1
Train classifier
C1
Classify all data objects
1 2´ 3 4´ 5 6´ 7 8 9 10
Update weights
.07 .17 .07 .17 .07 .17 .07 .07 .07 .07
New training data set
6 4 7 3 2 4 10 2 5 6
Train classifier
C2

AdaBoost

Algorithm 6: AdaBoost algorithm

1: Initialize $w_i(1) = \frac{1}{N}$ for $i = 1, \ldots, N$
2: **for** $t = 1, \ldots, T$ **do**
3: Create $\mathcal{D}_t$ by sampling (with replacement) from $\mathcal{D}$ according to $\boldsymbol{w}(t)$
4: Let f_t be the classifier *trained* on $\mathcal{D}_t$
5: $\epsilon_t = \sum_{i=1}^{N} w_i(t) \left(1 - \delta_{f_t(\boldsymbol{x}_i), y_i}\right)$ *(weighted error of f_t on all data).*
6: $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
7: For each i update weights using eq. (15.7):

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^{N} \tilde{w}_j(t+1)}, \quad \tilde{w}_i(t+1) = \begin{cases} w_i(t) e^{-\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) = y_i \\ w_i(t) e^{\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) \neq y_i. \end{cases}$$

8: **end for**
9: $f^*(\boldsymbol{x}) = \arg\max_{y=1,2} \sum_{t=1}^{T} \alpha_t \delta_{f_t(\boldsymbol{x}), y}$ *(Majority voting classifier)*

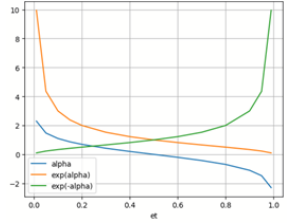
AdaBoost

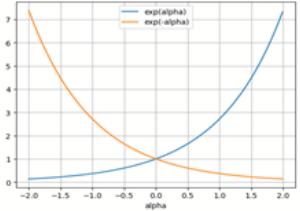
Algorithm 6: AdaBoost algorithm

1: Initialize $w_i(1) = \frac{1}{N}$ for $i = 1, \ldots, N$
2: **for** $t = 1, \ldots, T$ **do**
3: Create $\mathcal{D}_t$ by sampling (with replacement) from $\mathcal{D}$ according to $\boldsymbol{w}(t)$
4: Let f_t be the classifier *trained* on $\mathcal{D}_t$
5: $\epsilon_t = \sum_{i=1}^{N} w_i(t) \left(1 - \delta_{f_t(\boldsymbol{x}_i), y_i}\right)$ *(weighted error of f_t on all data).*
6: $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
7: For each i update weights using eq. (15.7):

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^{N} \tilde{w}_j(t+1)}, \quad \tilde{w}_i(t+1) = \begin{cases} w_i(t)e^{-\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) = y_i \\ w_i(t)e^{\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) \neq y_i. \end{cases}$$

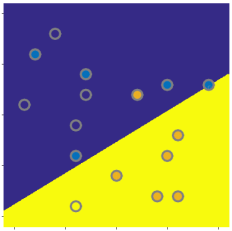
8: **end for**
9: $f^*(\boldsymbol{x}) = \arg\max_{y=1,2} \sum_{t=1}^{T} \alpha_t \delta_{f_t(\boldsymbol{x}), y}$ *(Majority voting classifier)*



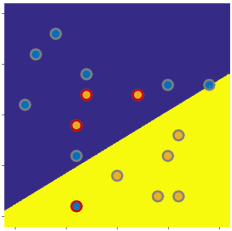


AdaBoost

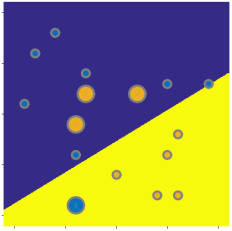
A:
A dataset is sampled with replacement and a classifier trained.



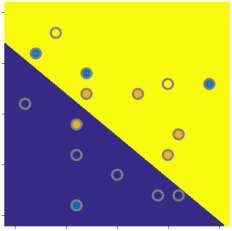
B:
Mis-classified observations are identified.



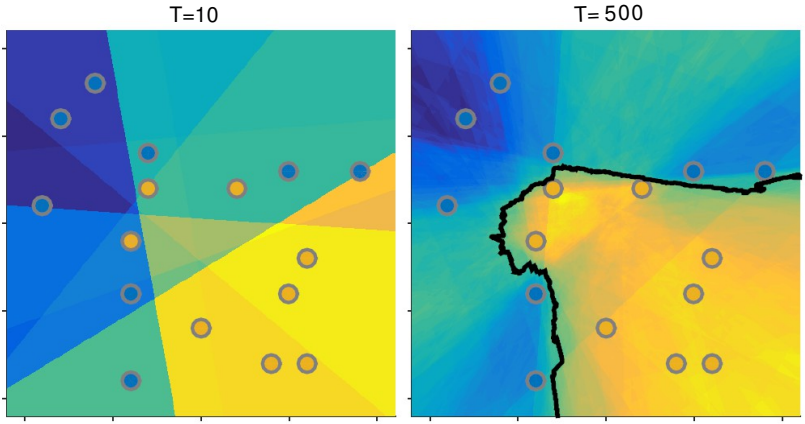
C:
Weighs are updated such that more emphasis is given to these mis-classified observations.



New round:
Based on the updated weights a new dataset is sampled and a classifier trained (shown), mis-classified observations identified and given more emphasis...



AdaBoost

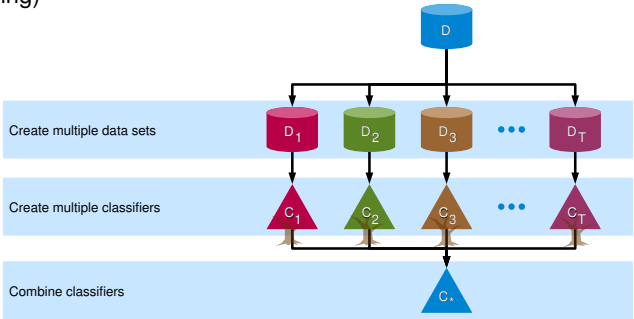


Bagging example: Random forest

Each tree is generated as follows:

- Sample dataset with replacement
- When generating each node in the tree, randomly select a subset of the features and only consider splits using these features

A large number of trees are generated and the trees are combined using majority voting (bagging)



Quiz 1: Adaboost

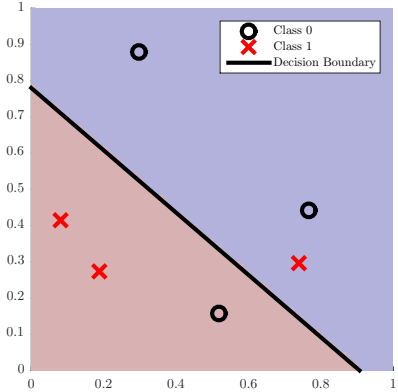


Figure 1: A binary classification problem and the decision boundary obtained by logistic regression. Observations left of the boundary are classified as belonging to the positive class 1 (red crosses) and observations right of the boundary to the negative class 0 (black circles)

We wish to apply a logistic regression model to the binary classification problem shown in Figure 1. We attempt to improve the performance by applying AdaBoost. AdaBoost works by first sampling a new dataset with replacement, then training a classifier on the dataset and then proceeding with the subsequent steps of the AdaBoost algorithm.

Suppose in the first iteration of the AdaBoost algorithm the classification boundary of the trained classifier is as indicated by the black line (i.e. observations left of the black line are classified as in the positive class). What is the resulting value of the weights $\boldsymbol{w}$?

A. $\boldsymbol{w} = \begin{bmatrix} 0.125 & 0.250 & 0.125 & 0.125 & 0.125 & 0.250 \end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix} 0.026 & 0.447 & 0.026 & 0.026 & 0.026 & 0.447 \end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix} 0.235 & 0.029 & 0.235 & 0.235 & 0.235 & 0.029 \end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix} 0.1 & 0.3 & 0.1 & 0.1 & 0.1 & 0.3 \end{bmatrix}$

E. Don't know.

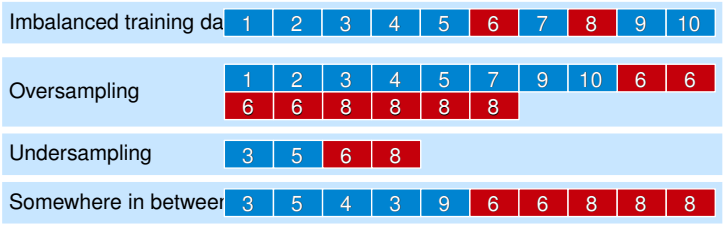
(Hint: First compute ε_1, then α_1, then the weights)

Class imbalance problem

- Many data sets have **imbalanced class distributions**
  - Example: Detection of defects that only occur rarely (e.g. 1/1,000,000)
  - Danger: Algorithm that says nothing is defect will be 99.999% correct
- **Solution approaches**
  - Resample to balance data sets
  - Modify existing classification algorithms
  - Measure performance in a way that takes balance into account

Resampling balanced data

- New sample has equal number of data objects from each class
- Approaches:
  - Undersampling: majority class: Throws out potentially useful data
  - Oversampling minority class: Increase data size and computational burden
  - Somewhere in between...



Confusion matrix

		Predicted: *positive*	Predicted: *negative*
Actual	*positive*	TP True Positive	FN False Negative
Actual	*negative*	FP False Positive	TN True Negative

Precision and recall

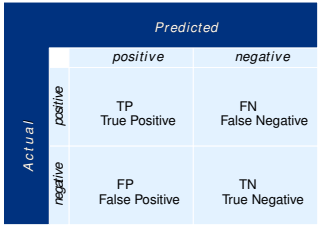
- **Precision**
  - Fraction of true positive among objects predicted to be positive
    $p = \frac{TP}{TP+FP}$
- **Recall**
  - Fraction of objects predicted to be positive among all positive objects
    $r = \frac{TP}{TP+FN}$

		Predicted: positive	Predicted: negative
Actual	positive	TP True Positive	FN False Negative
Actual	negative	FP False Positive	TN True Negative

Quiz 2: Precision/Recall

Consider two different classifiers, and suppose on a test set with 20 positive observations:

- Classifier 1 detects 54 positive of which 18 are actually positive
- Classifier 2 detects 16 positive of which 14 are actually positive



What is the precision and recall of the two classifiers?

A. Classifier 1: $p_1 = \frac{2}{5}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{1}{5}$

B. Classifier 1: $p_1 = \frac{1}{3}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{2}{3}, r_2 = \frac{9}{10}$

C. Classifier 1: $p_1 = \frac{2}{5}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{9}{10}$

D. Classifier 1: $p_1 = \frac{1}{3}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{7}{8}, r_2 = \frac{7}{10}$

Which classifier would you use if the objective was to detect credit-card fraud (the positive class corresponds to fraud)

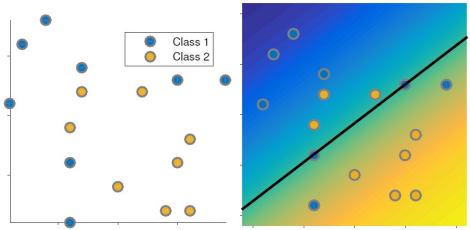
- **Precision**
  - Fraction of true positive among objects predicted to be positive

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- **Recall**
  - Fraction of objects predicted to be positive among all positive objects

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

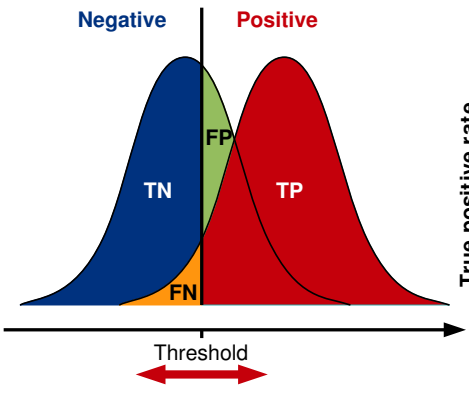
Classification threshold

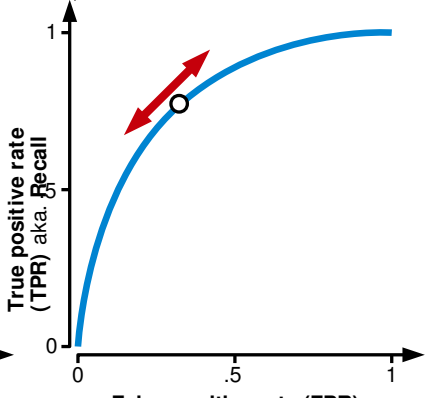


		Predicted: positive	Predicted: negative
Actual	positive	TP True Positive	FN False Negative
Actual	negative	FP False Positive	TN True Negative

Receiver operating characteristic (ROC)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



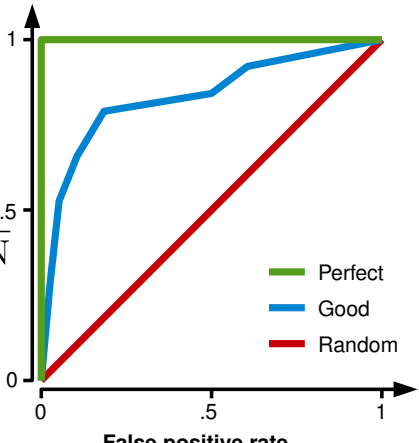


$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

Receiver operating characteristic (ROC)

True positive rate
aka. **Recall**

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

Quiz 3: AUC

	3 gears ($x_5 = 3$)	4 gears ($x_5 = 4$)	5 gears ($x_5 = 5$)
Low mpg ($y = 0$)	13	2	2
High mpg ($y = 1$)	2	10	3

Table 1: Number of low mpg and high mpg cars (i.e. $y = 0$ and $y = 1$) according to the number of gears, i.e. $x_5 = 3$, $x_5 = 4$, or $x_5 = 5$.

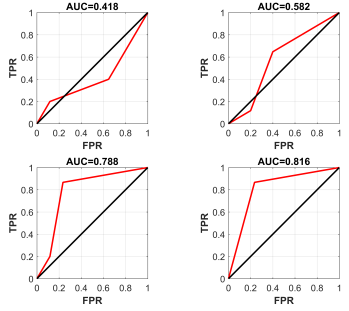


Figure 1: Four different receiver operator characteristic (ROC) curves and their area under curve (AUC) value.

A dataset representing cars contain an attribute x_5 corresponding to the number of gears. We wish to evaluate how well the number of gears predict low mpg, ($y = 0$, considered the negative class) from high mpg, ($y = 1$, considered the positive class) based on the data given in Table 1. For this purpose, we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature x_5. Which one of the ROC curves given in Figure 1 corresponds to using x_5 to discriminate between low mpg ($y = 0$) and high mpg ($y = 1$)?

A. The curve having AUC=0.418

B. The curve having AUC=0.582

C. The curve having AUC=0.788

D. The curve having AUC=0.816

E. Don't know.

(Hint: Select a value e.g. $x_5 = 4$. We then predict cars with 4 or more gears as being in the positive class and otherwise negative. Compute the FPR and TPR using this prediction and use the (FPR, TPR) values to discriminate between the curves)

Summary

- Ensemble methods
  - Bagging (incl. random forrest)
  - Boosting
- Evaluation
  - Imbalanced dataset
  - Confusion matrix
  - Recall, precision, TPR, FPR
  - ROC and AUC

Resources

https://www.youtube.com Video tutorial on ROC curve and AUC (https://www.youtube.com/watch?v=OAl6eAyP-yo)

https://towardsdatascience.com More in-depth discussion of the Random Forrest algorithm and parameter choices (https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd)

https://www.datacamp.com Practical use of the random forrest algorithm in python (https://www.datacamp.com/community/tutorials/random-forests-classifier-python)

https://citeseerx.ist.psu.edu Justification for the AdaBoost algorithm (technical) (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.9525)

Mid-term quiz 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

A. YEAR is ratio, SIZE is interval and LOCATION is nominal

B. YEAR is interval, SIZE is ratio and LOCATION is nominal

C. YEAR is interval, SIZE is ratio and LOCATION is ordinal

D. YEAR is interval, SIZE is ratio and LOCATION is interval

E. Don't know.

Year data types do not have a zero with a physical meaning and are therefore interval. Size has a physically relevant zero and is therefore ratio. Meanwhile, location is just an identifier which only support similarity-comparison and is therefore nominal.

Mid-term quiz 2

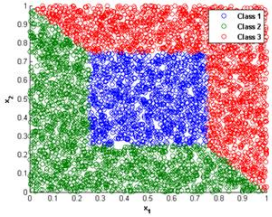
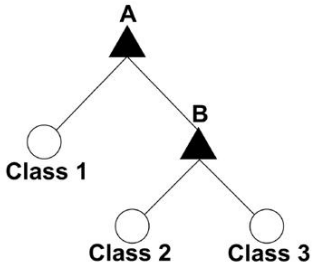


Figure 1



Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B. We will let $\boldsymbol{x}_n = (x, y)$ denote a 2-dimensional observation such that $\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of $\boldsymbol{x}_n$. Which one of the following classification rules would lead to a correct classification of the data?

A. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

B. A: $\|\boldsymbol{x}_n\|_1 \leq 1$, B: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$

C. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

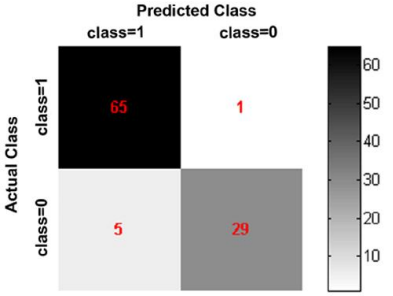
D. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B : $\|\boldsymbol{x}_n\|_1 \leq 1$

E. Don't know.

The right answer is D. Recall the shape associated with the different L_p-norms: $p = 2$ is a circle (Euclidean distance), $p = \infty$ a square, and $p = 1$ a square rotated 45 degrees. If we therefore first as at A if the observation is within the square (if yes, classify as the blue class, otherwise go to next split) and then at the next split ask if it within the L_1-norm of origo (distance 1), we get the diagonal decision boundary. This can be implemented by option D.

Mid-term quiz 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



A. The Accuracy is 94% and the Error rate is 6%

B. The Accuracy is 6% and the Error rate is 94%

C. The Accuracy is 65% and the Error rate is 35%

D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.

E. Don't know.

Accuracy is total number of correct choices divided by total number of observations. Therefore, the accuracy is $\frac{65+29}{6+65+29} = \frac{94}{100}$ or 95%. The right answer is therefore A.

Mid-term quiz 4

Which statement about crossvalidation is wrong?

A. Cross-validation can be used to estimate the generalization error.

B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.

C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.

D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.

E. Don't know.

The last option D is wrong because if we both select a model on a test set and then later use it for estimating the generalization error we will not obtain an unbiased estimate of the generalization error since we have already tuned the model on the test set. For this task, one should use two-layer CV.

Mid-term quiz 5

Consider a data set of four features: A, B, C, and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

A. C

B. B and C and D

C. A and B

D. A and B and C

E. Don't know.

Forward selection will attempt to minimize the error rate. It will first select C, then select lowest of the next options containing C, i.e. A, C, and then A, B, C. Therefore, option D is correct.

Mid-term quiz 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i[p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t. We will use Huntâ€™s algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{ Parent }) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j. We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Viriginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

A. The purity gain is $\Delta = \frac{3}{5}$

B. The purity gain is $\Delta = \frac{3}{15}$

C. The purity gain is $\Delta = \frac{6}{25}$

D. The purity gain is $\Delta = \frac{7}{15}$

E. Don't know.

There are a total of $N = 25$ observations and the number in the two branches are $N_1 = 10$ and $N_2 = 15$. In the base branch, the maximum class-probability is $\frac{10}{25}$ and so $I_0 = 1 - \frac{10}{25} = \frac{15}{25} = \frac{3}{5}$. Similarly, we compute $I_1 = \frac{1}{5}$ and $I_2 = 1 - \frac{8}{15} = \frac{7}{15}$. We now have

$$\Delta = I_0 - \frac{N_1}{N} I_1 - \frac{N_2}{N} I_2 = \frac{3}{5} - \frac{10}{25}\frac{1}{5} - \frac{15}{25}\frac{7}{15} \quad (1)$$

$$= \frac{3}{5} - 225 - \frac{7}{25} = \frac{15 - 2 - 7}{25} = \frac{6}{25} \quad (2)$$

or C.

Mid-term quiz 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing the test is well rested?

A. $\frac{4}{10}$

B. $\frac{8}{10}$

C. $\frac{9}{10}$

D. $\frac{10}{11}$

E. Don't know.

Let R be rested and P be passing. Then the answer is

$$P(R|P) = \frac{P(P|R)P(R)}{P(P|\overline{R})P(\overline{R}) + P(P|R)P(R)} \quad (1)$$

$$= \frac{0.9 \times 0.8}{0.4 \times 0.2 + 0.9 \times 0.8} \quad (2)$$

$$= \frac{0.9}{0.1 + 0.9} = \frac{9}{1 + 9} = \frac{9}{10} \quad (3)$$

and so C is correct.

Mid-term quiz 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

A. The first principal component accounts for more than 60% of the variation in the data.

B. The third principal component accounts for less than 5% of the variation in the data.

C. The second principal component accounts for more than 20% of the variation in the data.

D. The data can be perfectly represented in a three dimensional sub-space.

E. Don't know.

The variance explained of a given coordinate is $\frac{\sigma_i^2}{\sum_{i=1}^4 \sigma_i^2}$. Therefore, the variance explained by the second coordinate is $\frac{4}{21} < \frac{1}{5}$ and so C is the right answer.

Mid-term quiz 9

Consider the following sequence of numbers

$$x = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 3 & 4 & 4 & 5 & 14 \end{bmatrix}.$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y = \text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

A. $y = 1$

B. $y = 6$

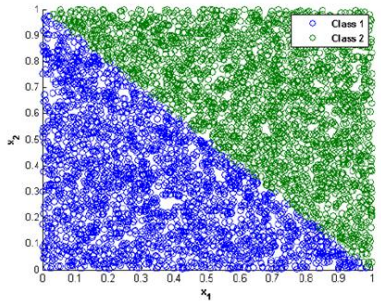
C. $y = 7$

D. $y = 11$

E. Don't know.

The mode is 1 (most common number). The median is 2.5 (since the list is ordered and contains an even number of elements it is the average of 2 and 3) and the mean is sum divided by 10 or 3.5. Therefore, the answer is 7.

Mid-term quiz 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w0. Which one of the following statements is **wrong**?

A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.

B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b, can perfectly separate the classes using only x_1 and x_2 as features.

C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.

D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.

E. Don't know.

B: To see why B is wrong, note the decision boundary will of such a tree will consist of rectangles with axis-oriented sides. The other options are easily seen to be correct and for C, note that the boundary shown in the plot corresopnsd to $z > 1$.