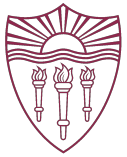


DSCI 531

SUPERVISED LEARNING

DECISION TREES + EXPLANATION

Kristina Lerman

USC Information Sciences Institute

Spring 2025

Notes on Quiz 1

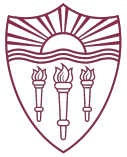
In-class

- You need to study what we discussed in lectures
- Requires pencil/pen
- Closed-book
- 10 minutes

Remotely (for DEN students only):

- Quiz PDF will be posted at the agreed-upon time
- You must have your camera with full view of your quiz and hands and screenshare on
- You have 5 minutes after the exam to send us a scan or picture of your quiz
- (you cannot write answers or make notes on screen as everyone will see what you write)

Reminders

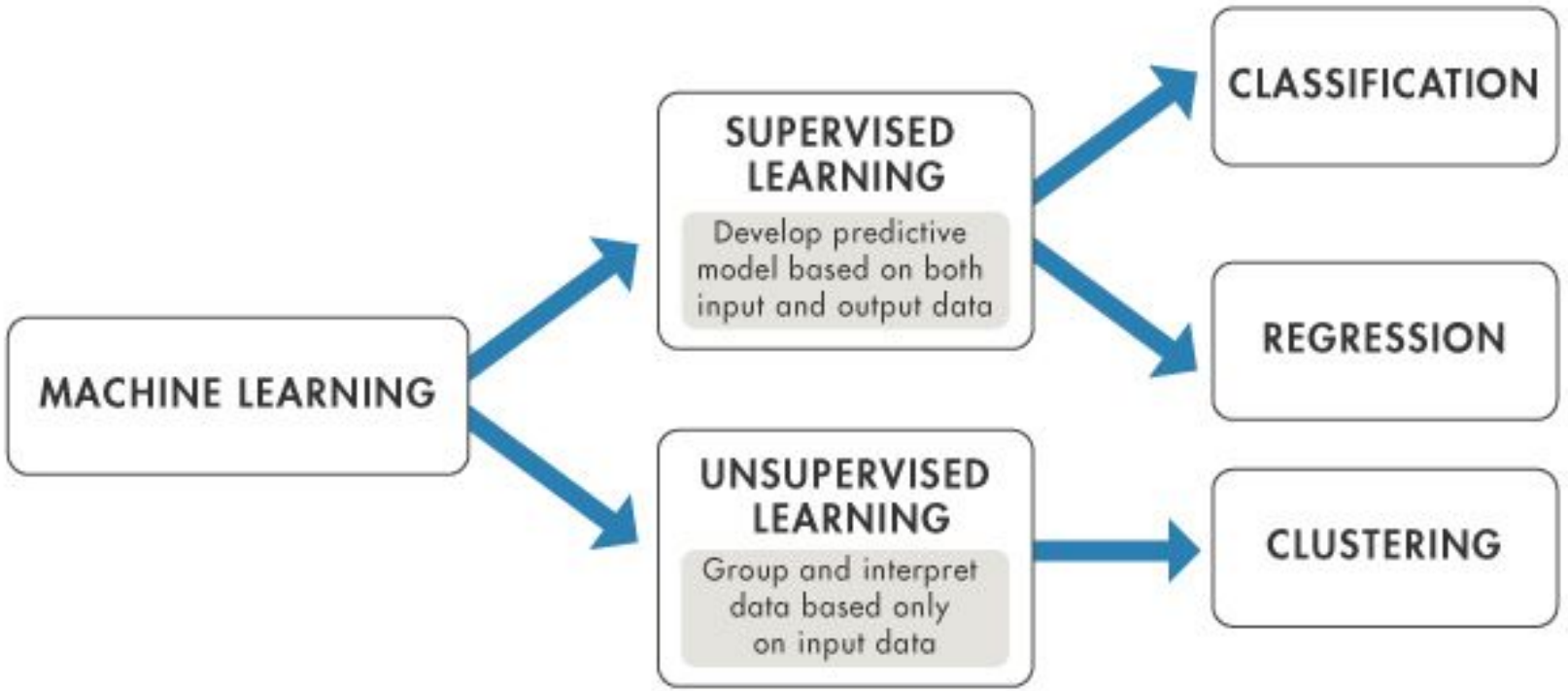


Homework 1 Due Jan 27th

Project Proposals Due Feb 5th

Supervised vs unsupervised learning

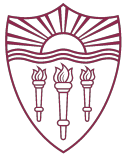
Based on sample/training data and their given class labels or categories, is it possible to train a model that generalizes over unseen data to decide what class the sample belongs to?



Source: DeepAI.org

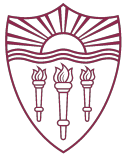
Output: **label**
- Binary
- Categorical

Output: **response/outcome**
- Binary
- Real



MIXED EFFECTS MODELS

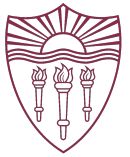
Independence assumption is often violated!



- More complex research questions require collecting multiple responses from the same subject
- But, multiple responses from the same subject cannot be regarded as independent from each other
  - Every person has a slightly different voice pitch, and this idiosyncratic factor will affect all responses from the same subject, making them inter-dependent rather than independent

subject	gender	scenario	frequency
F1	F	1	213.3
F1	F	1	204.5
F1	F	2	285.1
F1	F	2	259.7
F1	F	3	203.9
F1	F	3	286.9
F3	F	1	229.7
F3	F	1	237.3
F3	F	2	236.8
F3	F	2	251
F3	F	3	267
F3	F	3	266
M4	M	1	110.7
M4	M	1	123.6
M4	M	2	229
M4	M	2	114.9
M4	M	3	112.2

Independence assumption is often violated



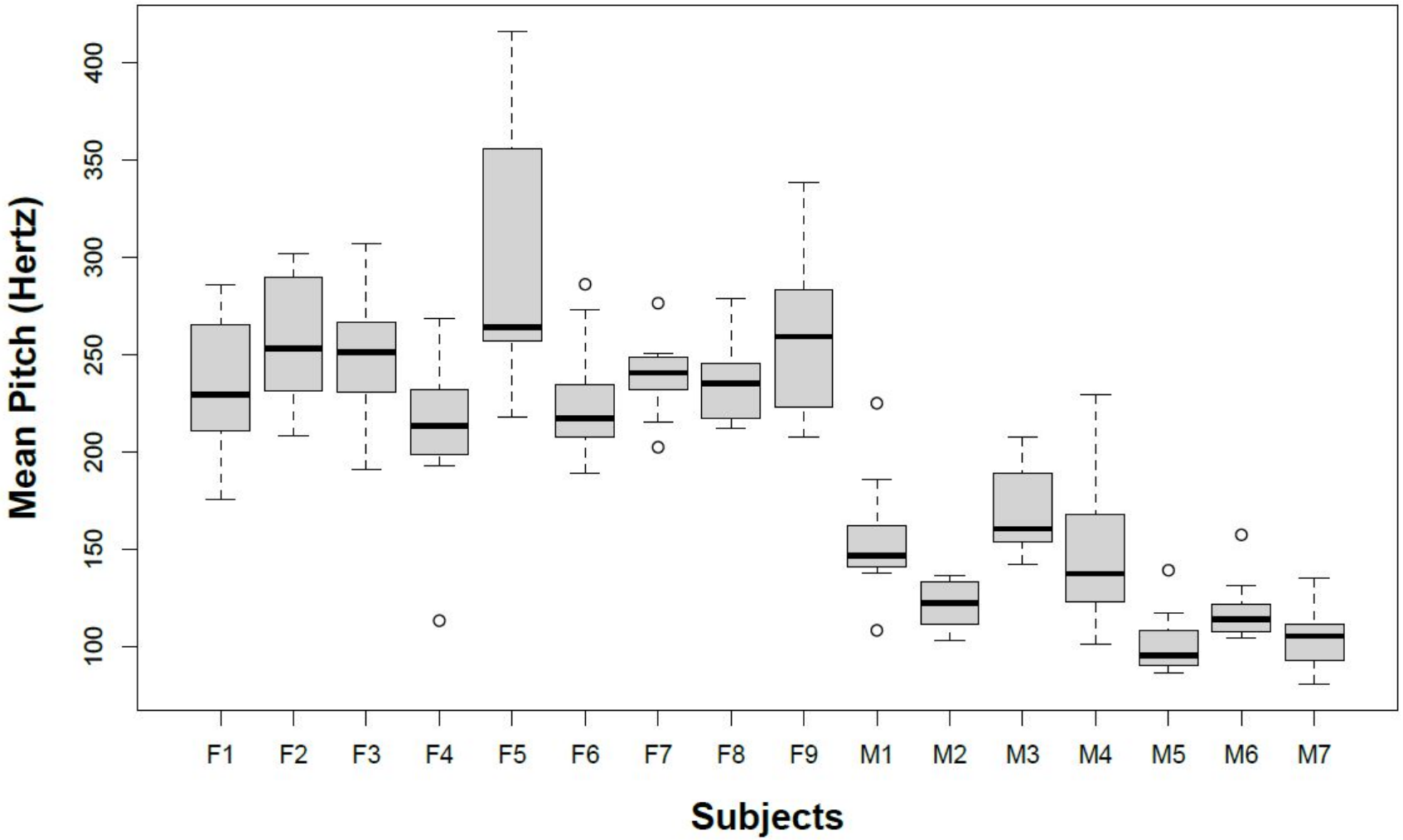
- **Research question: Does politeness affect pitch**?

  pitch ~ politeness + sex + ε

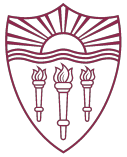
- Each subject gives multiple responses: **polite** and **informal** response
- Multiple responses from the same subject cannot be regarded as independent from each other
  - Every person has a slightly different voice pitch, and this idiosyncratic factor will affect all responses from the same subject, making them inter-dependent rather than independent
- Add a random effect
  - This allows us to resolve this non-independence by assuming a different "baseline" pitch for each subject.

Lots of individual variation

mixed_effect_on_politness_data.ipynb



Modeling individual differences with random effects



- Model individual differences by assuming different ***random intercepts*** for each subject.
  - Each subject is assigned a different intercept value, and the mixed model estimates these intercepts.
- Mixed model adds one or more **random effects** to the fixed effects model.
  - These random effects give structure to the error term "ε".
  - In the model, each "subject" becomes a random effect, and this characterizes idiosyncratic variation that is due to individual differences.

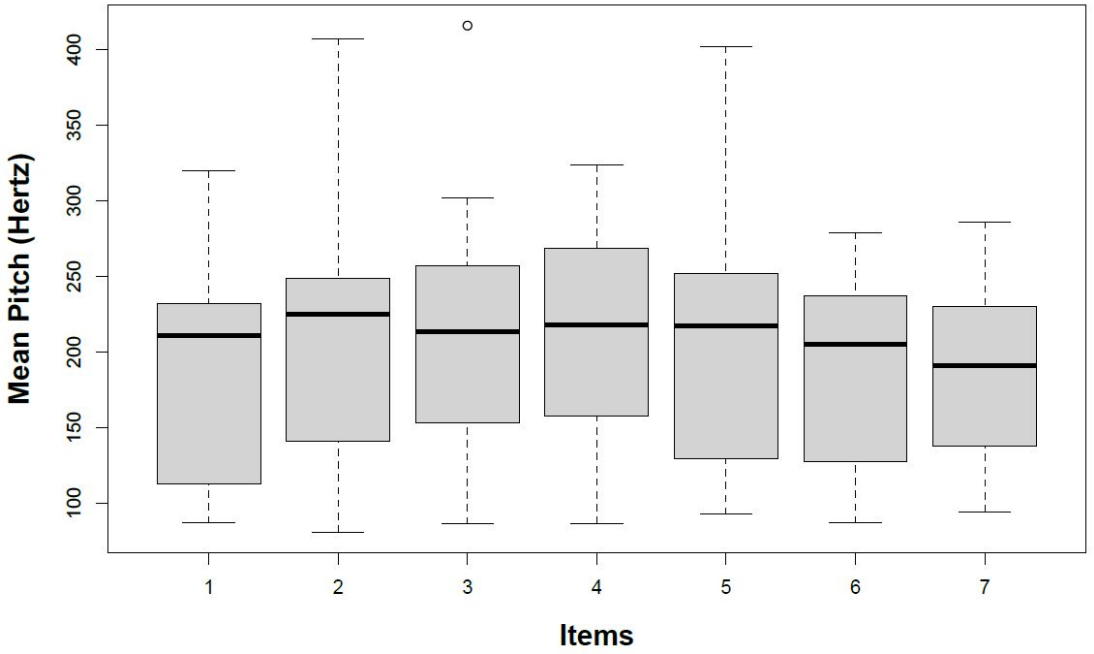
Mixed effects model

pitch ~ politeness + sex + (1|subject) + ε

- "(1|subject)" means a different intercept for each subject"
  - "1" stands for the intercept.
  - Formula tells the model to expect multiple responses per subject, and these responses will depend on each subject's baseline level.
  - This resolves the non-independence that stems from having multiple responses by the same subject.
- Error term "ε" captures remaining "random" differences between different utterances from the same subject.

Modeling multiple dependencies

- Systematic per-item variation
  - Some utterances (items) may have a higher pitch not explained by politeness and subject, but due to another factor that affects the voice pitch of all subjects (e.g., embarrassment)
  - Not accounting for this, violates the independence assumption



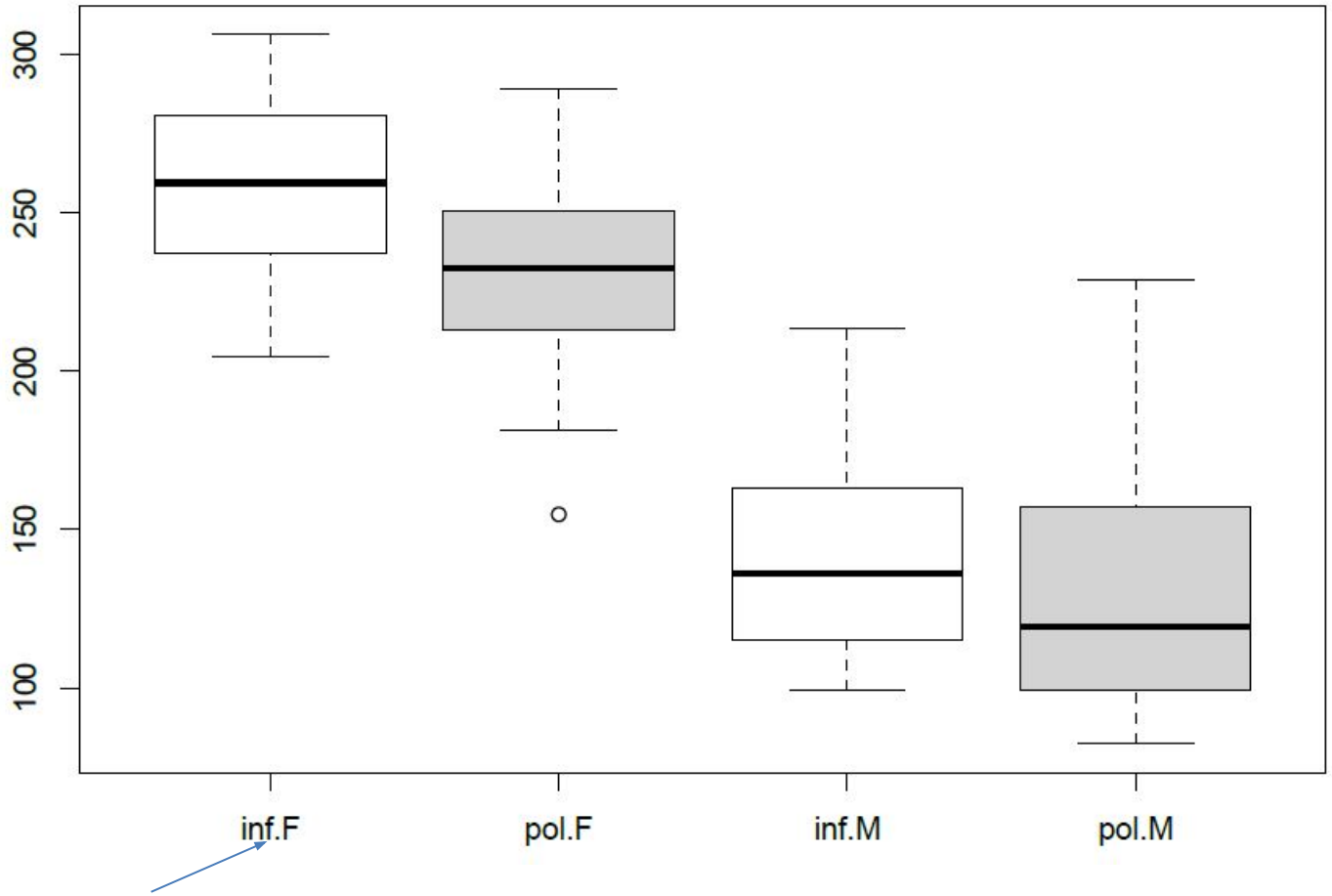
Multiple mixed effects model

$$\text{pitch} \sim \text{politeness} + \text{sex} + (1|\text{subject}) + (1|\text{item}) + \varepsilon$$

- The model knows there are multiple responses per subject and per item
  - 1|subject : Different intercepts for different subjects
  - 1|item: Different intercepts for different items.
- We now “resolved” these non-independencies and accounted for per-subject and per-item variation in overall pitch levels.

Illustration on “Politeness” data

pitch ~ attitude + sex + (1|subject) + (1|item) + ε



Informal speech by females

Polite speech by males

Random effects

pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

```
Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 205.2    14.33
 subject  (Intercept) 417.0    20.42
 Residual             637.4    25.25
```

- Gender explains much of the between-subject variability in pitch. Without explicitly modeling gender, the subject variance is much higher.
- Residual - ε term – is the variability that is not due to “item” or “subject”

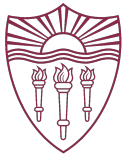
Fixed effects

pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  256.847     13.827  18.576
attitudepol  -19.722      5.547  -3.555
genderM     -108.517     17.572  -6.176
```

- The coefficient “attitudepol” is the slope for the categorical effect of politeness.
  - Minus 19.695 means going from “informal” to “polite” utterances decreases the pitch by -19.695 Hz.
  - Polite speech has lower pitch than informal speech
- Coefficient of “genderM” is negative
  - Males have lower pitch than females

Statistical significance of mixed effects models



- Variety of opinions about the best approach
- Likelihood ratio test
  - Probability of observing the data you collected given the model you learned.
- The logic of the likelihood ratio test is to compare the likelihood of two models with each other.
  - *Null model*: The model *without* the factor that you're interested in

    pitch ~ gender + (1|subject) + (1|scenario) + ε
  - *Full model*: *with* the factor that you're interested in.

    pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

Likelihood ratio test

```
Data: politeness
Models:
politeness.null: frequency ~ gender + (1 | subject) + (1 | scenario)
politeness.model: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
                 Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
politeness.null   5 816.72 828.81 -403.36   806.72
politeness.model  6 807.10 821.61 -397.55   795.10 11.618      1  0.0006532 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- perform the likelihood ratio test using a standard package (eg, anova)
- Report the effect in a paper as follows:
  - “… politeness affected pitch ($\chi 2(1)=11.62$, $p=0.00065$), lowering it by about 19.7 Hz ± 5.6 (standard errors) …”

Random slopes vs random intercepts

pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

- Different intercept for each subject and each item
  - Different baselines for each subject
- But the same coefficients: Fixed effects of gender and attitude are the same for all subjects and items
- Need a model with random slopes to allow subjects to have individualized responses to fixed effects

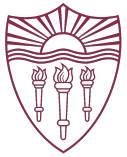
```
$scenario
  (Intercept) attitudepol   genderM
1    243.4859   -19.72207 -108.5173
2    263.3592   -19.72207 -108.5173
3    268.1322   -19.72207 -108.5173
4    277.2546   -19.72207 -108.5173
5    254.9319   -19.72207 -108.5173
6    244.8015   -19.72207 -108.5173
7    245.9618   -19.72207 -108.5173

$subject
   (Intercept) attitudepol   genderM
F1    243.3684   -19.72207 -108.5173
F2    266.9443   -19.72207 -108.5173
F3    260.2276   -19.72207 -108.5173
M3    284.3536   -19.72207 -108.5173
M4    262.0575   -19.72207 -108.5173
M7    224.1292   -19.72207 -108.5173

attr(,"class")
[1] "coef.mer"
```

Mixed effects with random slopes

pitch ~ attitude + gender + (1+attitude|subject) + (1+attitude|scenario) + ε

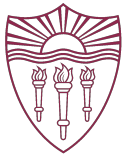
- Coefficient for the effect of politeness (“attitudepol”) is different for each subject and item
- despite individual variation, there is also consistency in how politeness affects voice: pitch tends to go down when speaking politely

```
$scenario
  (Intercept) attitudepol   genderM
1    245.2603   -20.43832 -110.8021
2    263.3012   -15.94386 -110.8021
3    269.1432   -20.63361 -110.8021
4    276.8309   -16.30132 -110.8021
5    256.0579   -19.40575 -110.8021
6    246.8605   -21.94816 -110.8021
7    248.4702   -23.55752 -110.8021

$subject
   (Intercept) attitudepol   genderM
F1    243.8053   -20.68245 -110.8021
F2    266.7321   -19.17028 -110.8021
F3    260.1484   -19.60452 -110.8021
M3    285.6958   -17.91950 -110.8021
M4    264.1982   -19.33741 -110.8021
M7    227.3551   -21.76744 -110.8021

attr(,"class")
[1] "coef.mer"
```



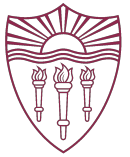
DECISION TREES

Why decision trees?



- Popular method, especially for classification
- Learn quickly, classify new data quickly
- Non-parameteric
- Non-linear
- Highly accurate
- Interpretable
  - Set of IF-THEN rules
  - Can be validated by human experts

Decision Trees

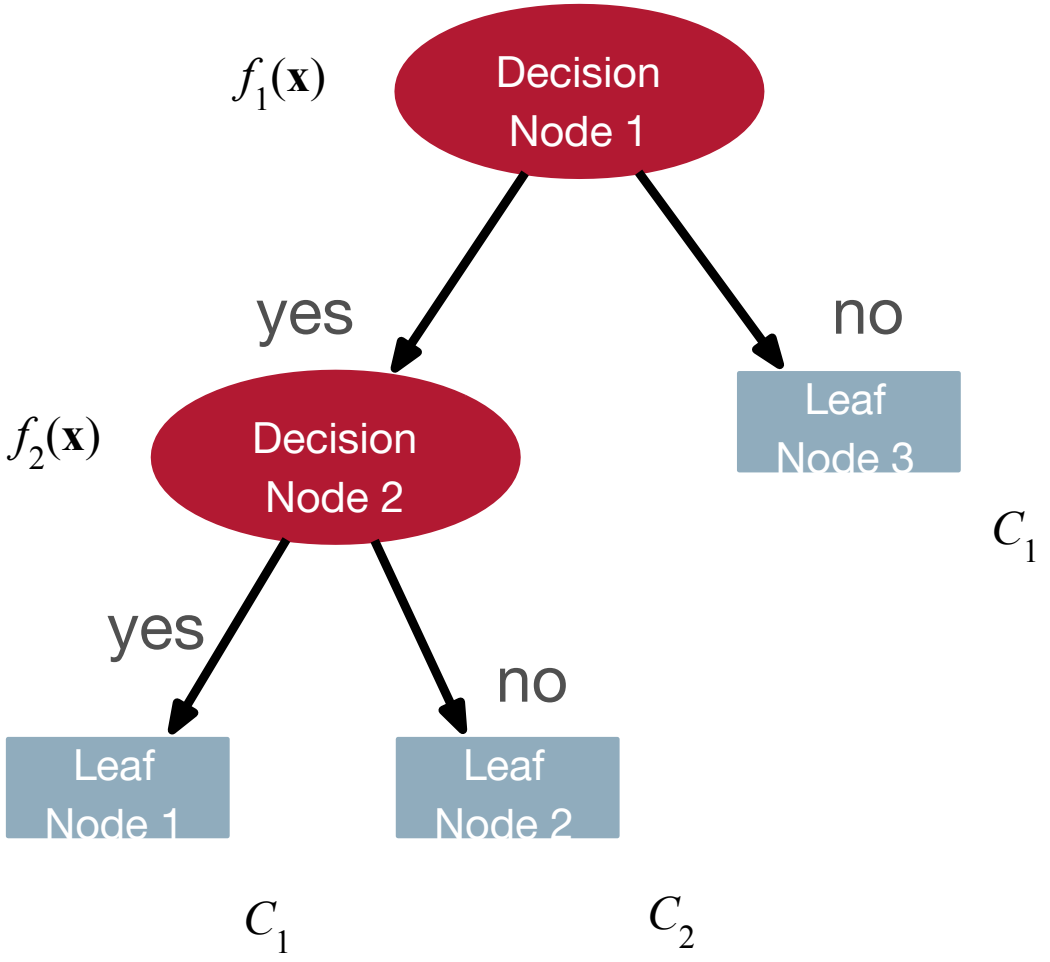


- A *decision tree* is a hierarchical data structure following the "*divide and conquer*" strategy
- Nonparametric ("no predefined parameters") method that is used for:
  - Classification
  - Regression
- Tree based algorithms involve *stratifying* or *segmenting* the *predictor space* into several simple *regions*.
- A set of splitting rules also called as *decisions* govern how this stratification is made.

- These splits of the feature space can be represented in a tree structure. Hence, the name – *decision trees.*

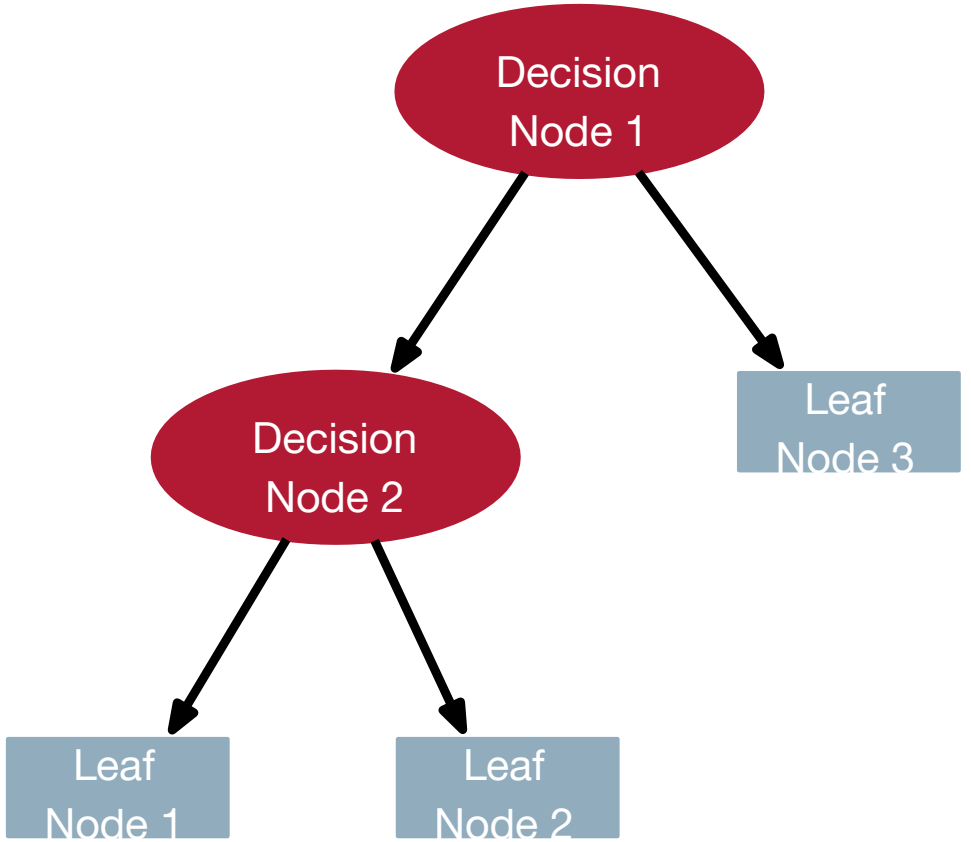
Decision Tree Structure

- A decision tree consists of:
  - Decision nodes
  - Leaf nodes □ data resides here
- Each decision node *m* implements a test function $f_m(\boldsymbol{x})$
- Depending on the test function outcome a branch in the tree is taken.
- This is applied recursively until a *leaf node* is reached.
- The value in the leaf node defines the output.

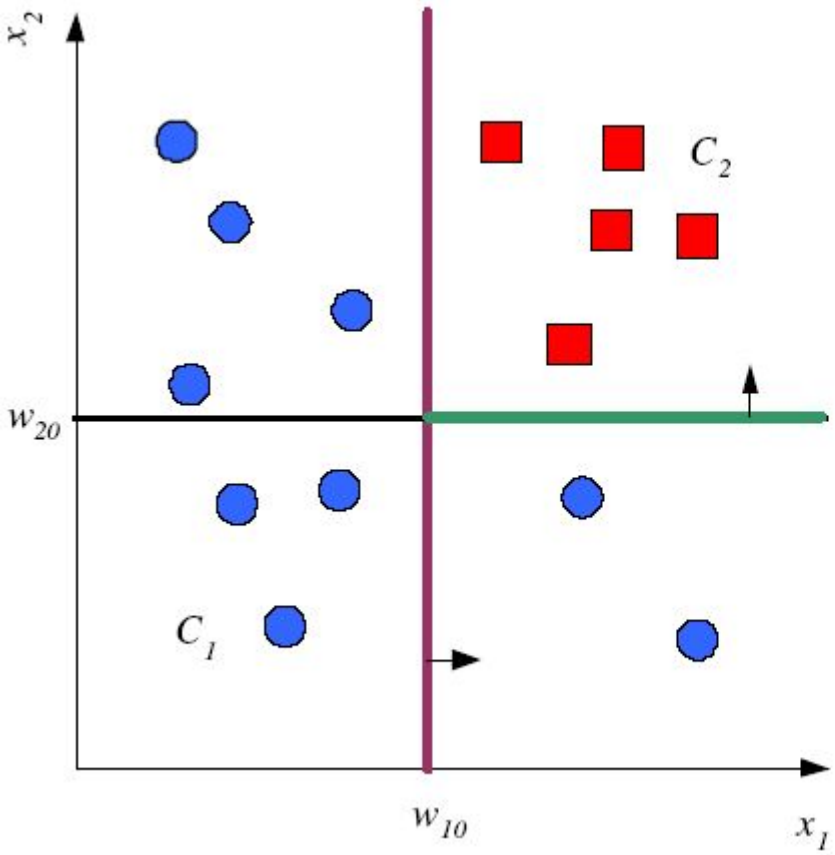


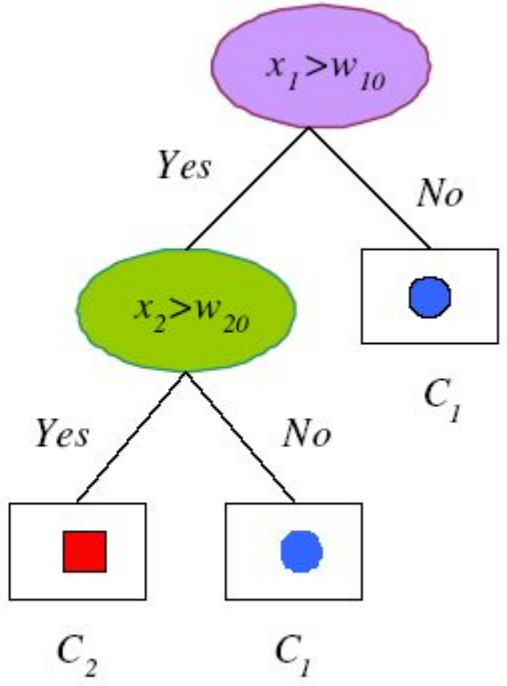
Decision Tree Characteristics

- It is a nonparametric method, as we do not know a priori how the tree will look like in the end. We allow it to grow branches and leaves depending on the complexity of the data.
- In addition, no assumptions about the class densities of data are assumed.



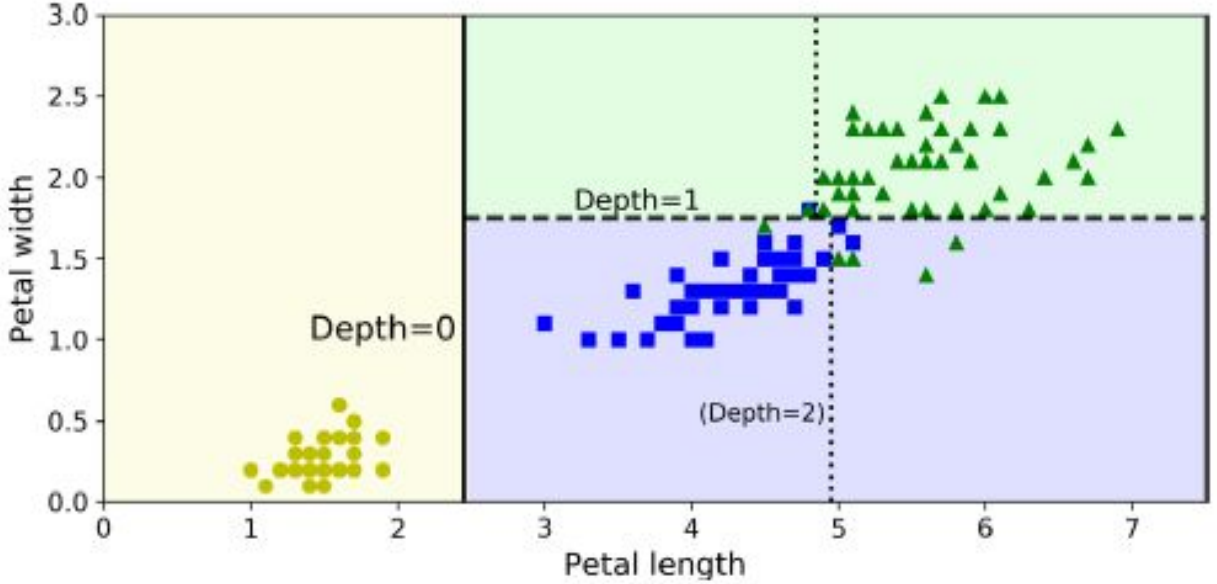
Decision Tree Example

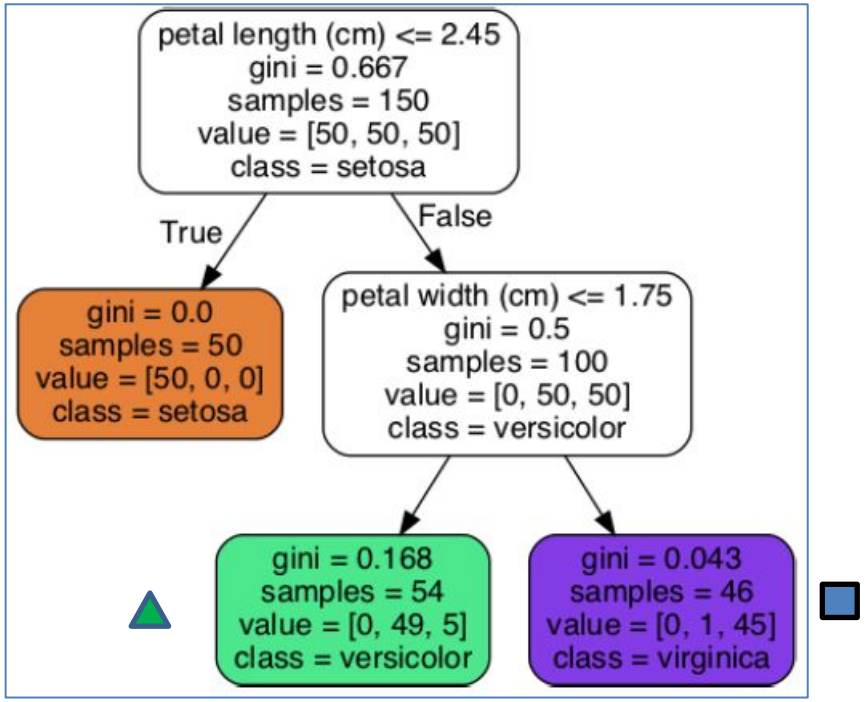


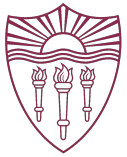


UCI Iris Data

Iris Versicolor **Iris Setosa** **Iris Virginica**

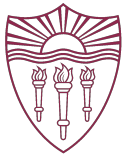






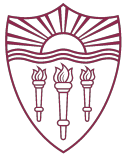
ENSEMBLE METHODS

Intuition



- No-Free-Lunch Theorem
  - There is no algorithm that is always the most accurate
- Objective
  - Generate a group of (less accurate) base-learners which when combined has higher accuracy
- Different learners use different
  - Algorithms
  - Hyperparameters
  - Representations /Modalities/Views
  - Training sets
  - Subproblems
- Diversity vs Accuracy

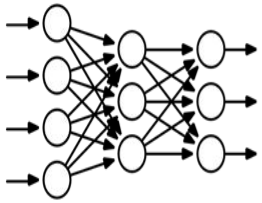
Diversity



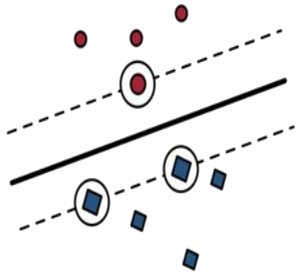
- *Analogy:* In order to learn from someone else it is important that the person does not have exactly the same knowledge as you do
- The same holds for classifiers: If two classifiers always make the same mistakes there is no way a combination of the two can improve performance
  - At the same time the classifiers or learners should be accurate in their own domain of expertise
- We have two goals
  - **maximize individual accuracies and**
  - **diversity between learners**

Achieving Diversity

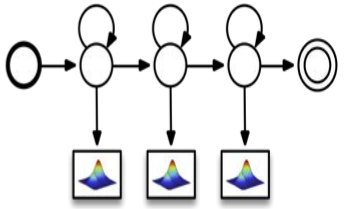
- Different Algorithms



Multi layer perceptron



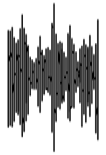
Support vector machine



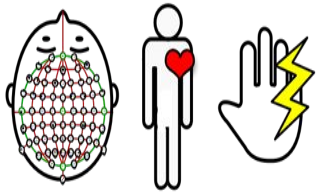
Hidden Markov model

- Different Input Representations

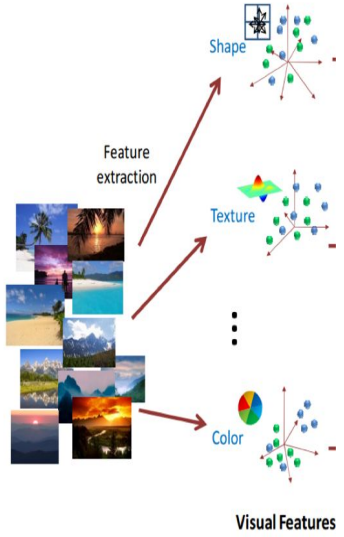
Video input



Audio waveform



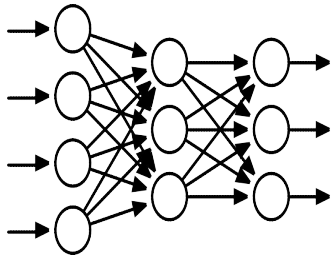
Physiological data



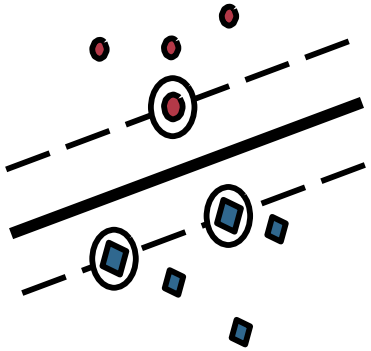
- Different Input Representations (Multi-view)
- Different Training Sets (boosting and bagging)

Different Algorithms

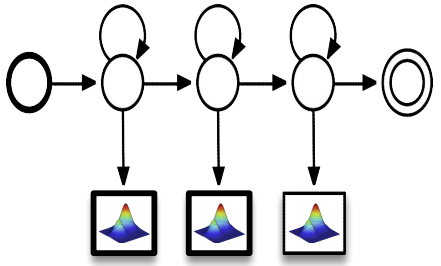
- Through the use of different base learners such as SVM, MLP, HMM we can enforce diversity to some degree
- E.g. an HMM would take temporal dynamics into account for the classification whereas the base models for SVM or MLP would not.



Multi layer perceptron

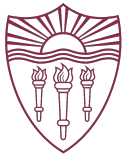


Support vector machine



Hidden Markov model

Different Hyper-Parameters

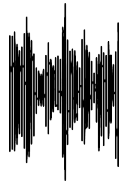


- We can use the same algorithms (e.g. multiple MLP) but use different hyper parameters for training.
- For example, different number of layers in the MLP or different number of neurons; different selection of kernel for SVM etc.
- For classifiers that utilize gradient descent methods even the initialization of weights that define the starting location of the learning algorithms can somewhat influence the outcome of the classifier. This diversity is likely not too strong though.

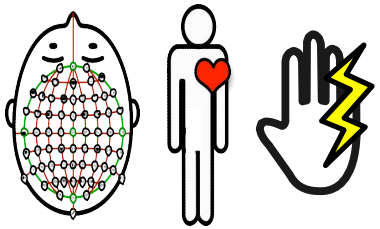
Different Input Representations

- Separate base learners may make use of different representations of the same input “object” or “event”.
- Information of different representations from different sources/sensors/modalities often lead to diversity in classifiers allowing for better identification.
  - In human emotion recognition or user state recognition many modalities contain information. We typically fuse audio, video, and sometimes physiology. This paradigm is used as **“sensor fusion” or “multimodal fusion”**

Video input

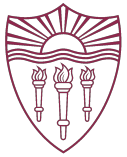


Audio waveform



Physiological data

Different Training Sets



- If sufficient data are available it is possible to train classifiers on different subsets of the data, which leads to different "views" on the same type of data
- This can be done through randomly choosing different subsets (named ***bagging***)
- Or the learners can be trained serially so that instances on which the previous classifiers are not accurate are given more *emphasis* in the training for later base learners (named ***boosting*** or ***cascading***)

Fusion Schema: Voting

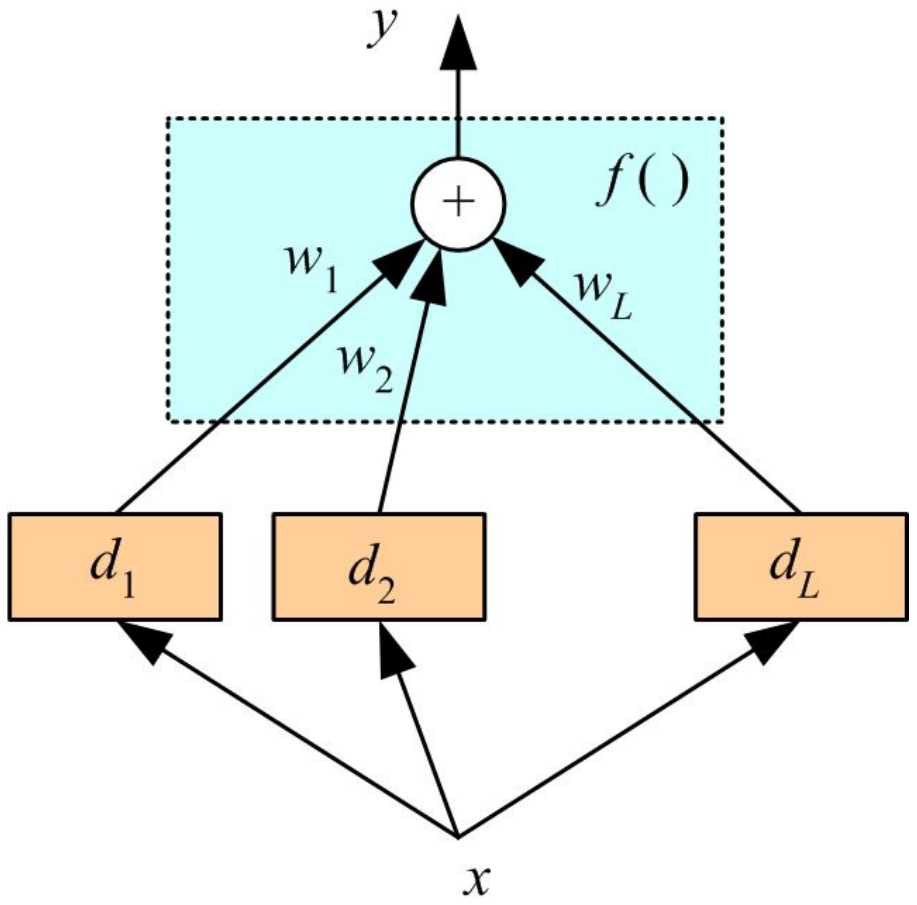
- Linear combination

$$y = \sum_{j=1}^{L} w_j d_j$$

$$w_j \geq 0 \text{ and } \sum_{j=1}^{L} w_j = 1$$

- Classification

$$y_i = \sum_{j=1}^{L} w_j d_{ji}$$

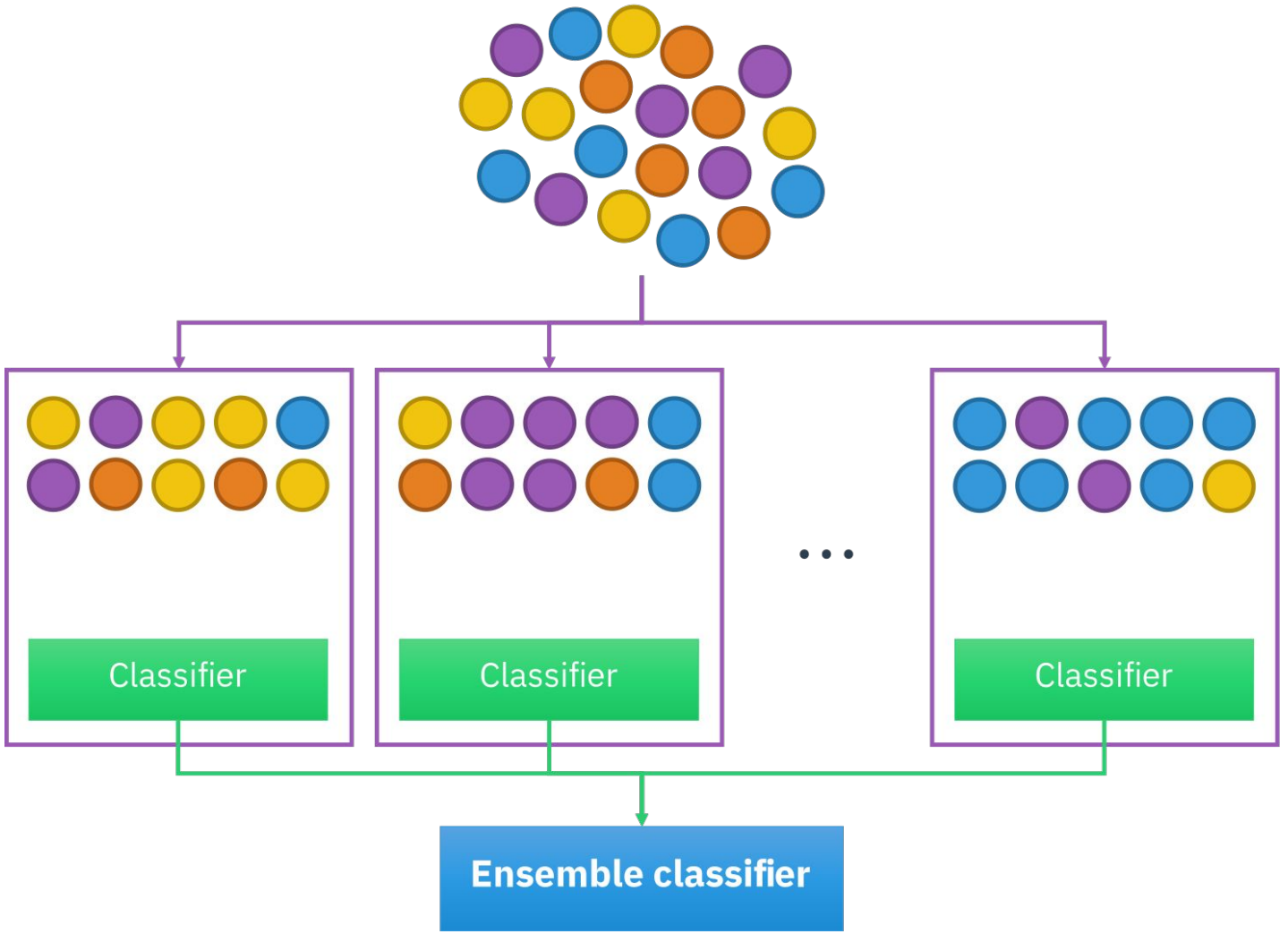


Fixed Combination Rules

Rule	Fusion function $f(\cdot)$
Sum	$y_i = \frac{1}{L}\sum_{j=1}^{L} d_{ji}$
Weighted sum	$y_i = \sum_j w_j d_{ji}, w_j \geq 0, \sum_j w_j = 1$
Median	$y_i = \text{median}_j d_{ji}$
Minimum	$y_i = \min_j d_{ji}$
Maximum	$y_i = \max_j d_{ji}$
Product	$y_i = \prod_j d_{ji}$

	C_1	C_2	C_3
d_1	0.2	0.5	0.3
d_2	0.0	0.6	0.4
d_3	0.4	0.4	0.2
Sum	0.2	**0.5**	0.3
Median	0.2	**0.5**	0.4
Minimum	0.0	**0.4**	0.2
Maximum	0.4	**0.6**	0.4
Product	0.0	**0.12**	0.032

Bagging (=bootstrap aggregation)



Original Data

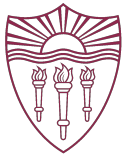
Bootstrapping

(sample data with replacement)

Aggregating

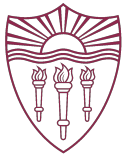
Use voting (Average or median with regression)

Bagging



RANDOM FOREST = ENSEMBLE OF DECISION TREES

Random Forest

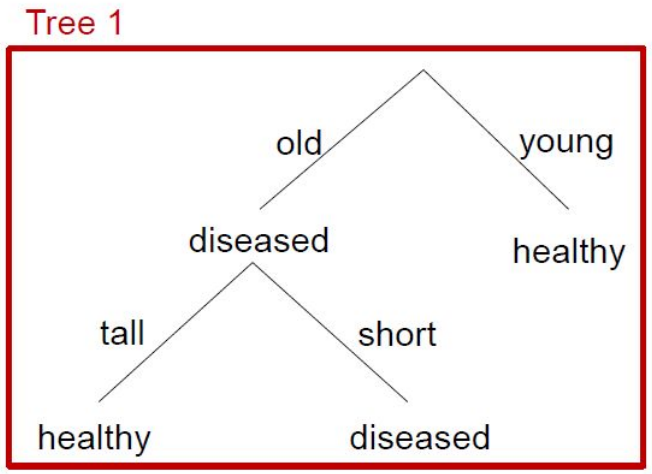


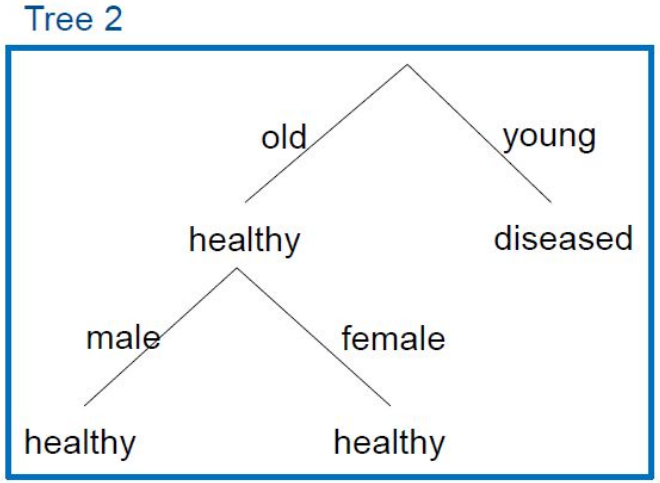
As in bagging, we build a number of decision trees on bootstrapped training samples each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of p predictors.

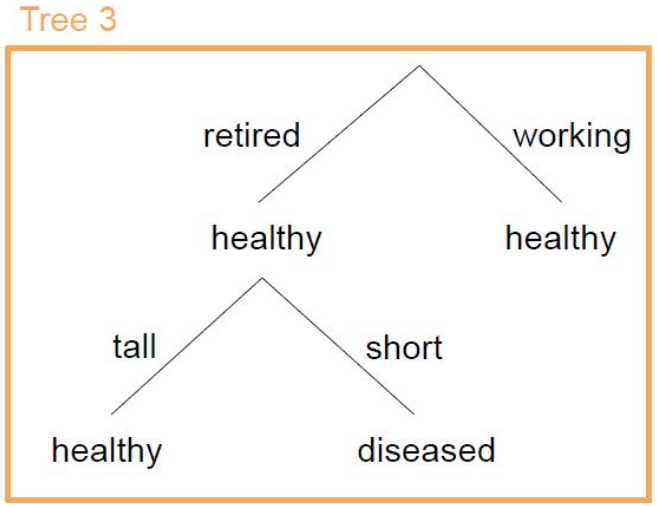
- Take a random sample (with replacement) of the training data

- To avoid creating highly correlated trees, choose a random sample (without replacement) of predictors (features).
  - De-correlation increases accuracy

Illustration



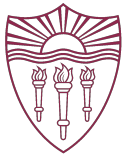




New sample: old, retired, male, short
Tree predictions: diseased, healthy, diseased

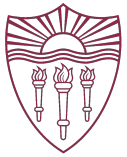
Majority rule:
diseased

Feature importance



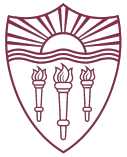
- measure feature importance by looking at how much the tree nodes that use that feature reduce impurity on average (across all trees in the forest).
- A weighted average, where each node's weight is equal to the number of training samples that are associated with the feature

Difference from decision trees



- Train each tree on bootstrap resample of data
  - Bootstrap resample of data set with N samples:
  - Make new data set by drawing with replacement N samples; i.e., some samples will probably occur multiple times in new data set
- For each split, consider only m randomly selected variables
- Don’t prune
- Fit B trees in such a way and use average or majority voting to aggregate results

Trees vs Random Forests

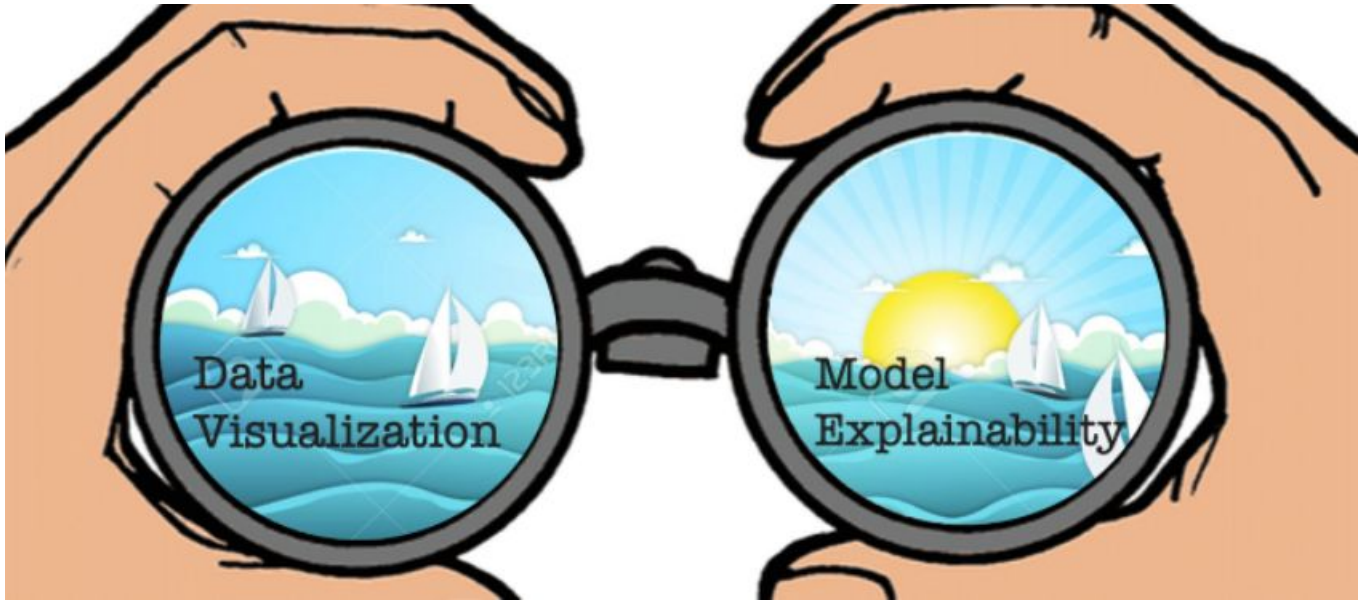


+ Trees yield insight into decision rules
+ Rather fast
+ Easy to tune parameters

- Prediction of trees tend to have a high variance

+ RF as smaller prediction variance and therefore usually a better general performance
+ Easy to tune parameters

- Rather slow
- “Black Box”: Rather hard to get insights into decision rules

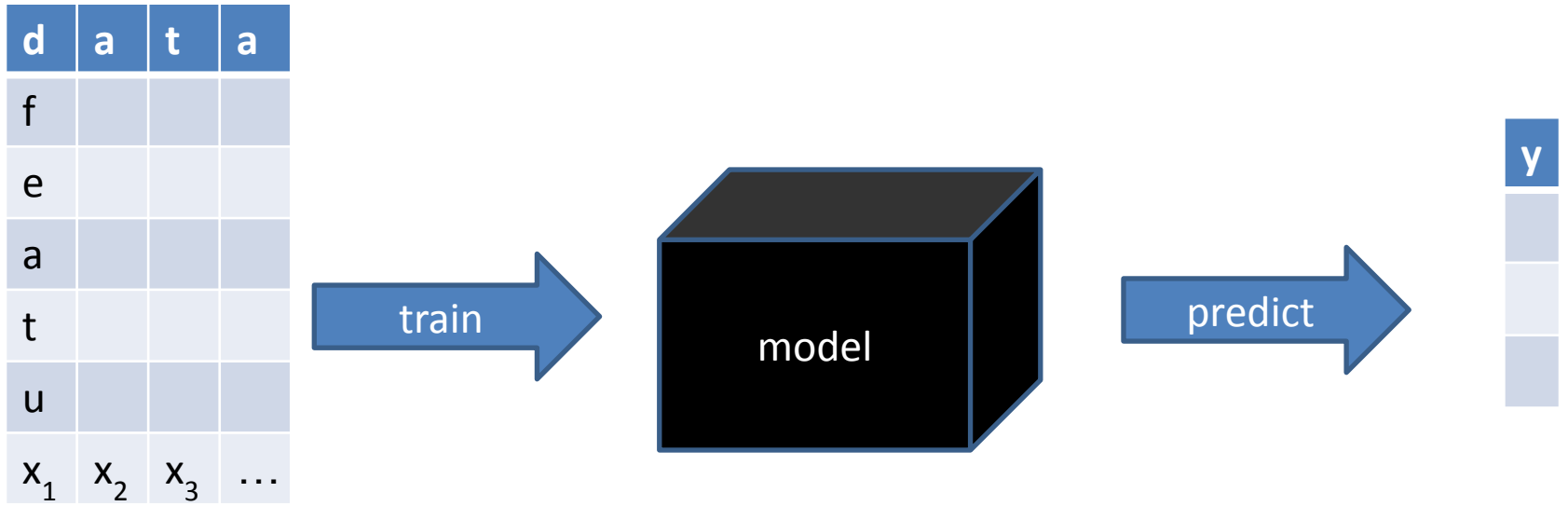


Source: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

SHAPLEY VALUES, S3D

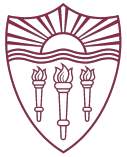
EXPLAINING THE DATA, EXPLAINING THE MODEL

Explaining the model with SHAP values



- What is the feature's contribution to the accuracy of predictions?
- **Shapley value: the average of the marginal contributions across all features.**

Benefits of SHAP values



- *global interpretability* — the collective SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable (cf feature importance).
- *local interpretability* — each observation gets its own set of SHAP values. This greatly increases its transparency. We can explain why a case receives its prediction and the contributions of the predictors. Unlike feature importance, local interpretability enables us to pinpoint and contrast the impacts of the factors on individual cases.
- SHAP values can be calculated for any tree-based model.

Illustration on wine quality data

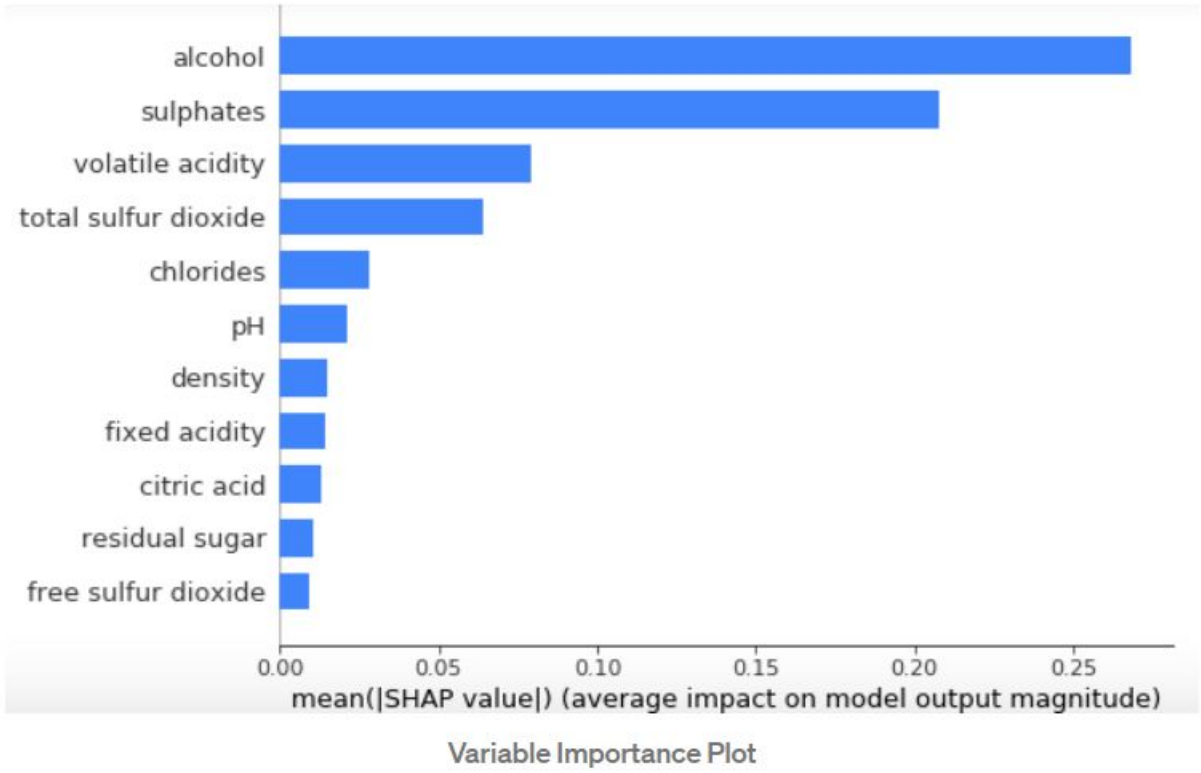
Outcome
- Wine quality

Dimensions
- Citric acid,
- Chlorides,
- Free Sulfur Dioxide (SO2),
- Residual sugars,
- …

data

	outcome :quality	free SO 2	citric acid	residual sugars	…
White wines	4898				
Red wines	1599				

Variable importance plot

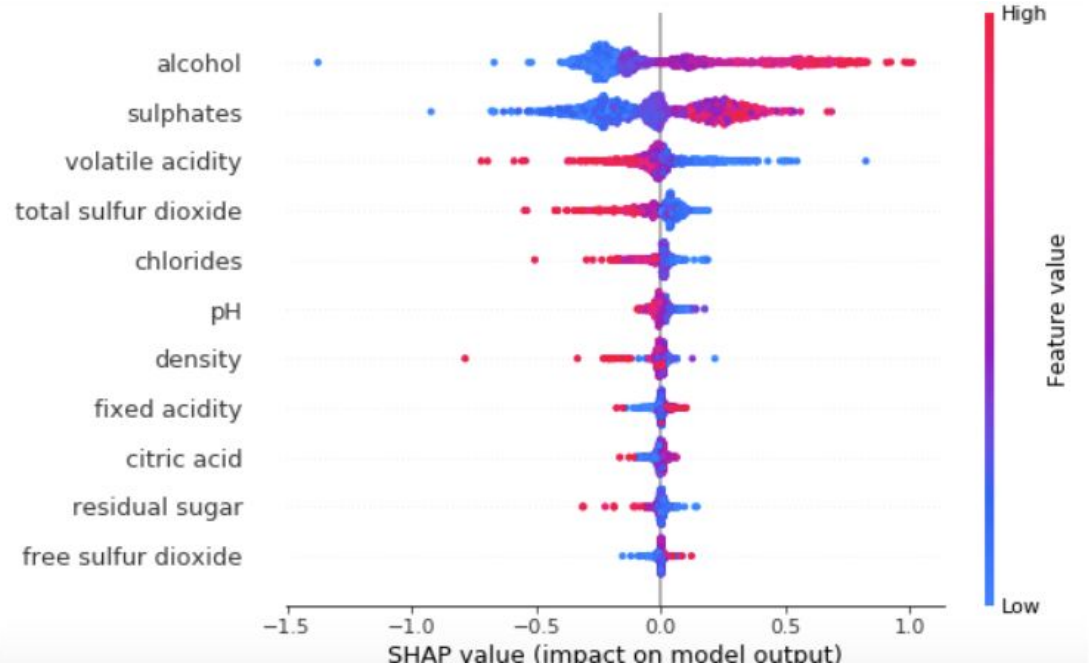


Variable Importance Plot

A variable importance plot lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power.

Variable importance can show positive and negative relationships with target variable

This plot shows all data points in the training data.



- *Feature importance:* Variables are ranked in descending order.
- *Impact:* The horizontal location shows whether the effect of that value *is associated with a higher or lower prediction*.
- *Original value:* Color shows whether that variable is high (in red) or low (in blue) for that observation.
- *Correlation:* A *high* level of the "alcohol" content has a high and *positive* impact on the quality rating. The "high" comes from the red color, and the "positive" impact is shown on the X-axis. Similarly, we will say the "volatile acidity" is negatively correlated with the target variable.

Local interpretability – each observation gets its own SHAP values

- The *output value* is the prediction for that observation.
- The *base value*: the value that would be predicted if we did not know any features for the current output.
- *Red/blue*: Features that push the prediction higher (to the right) are shown in red, and those pushing the prediction lower are in blue.
- *Alcohol*: has a positive impact on the quality rating. The alcohol content of this wine is 11.8 which is higher than the average value 10.41. So it pushes the prediction to the right.
- *pH*: has a negative impact on the quality rating. A lower than the average pH (=3.26 < 3.30) drives the prediction to the right.
- *Sulphates*: is positively related to the quality rating. A lower than the average Sulphates (= 0.64 < 0.65) pushes the prediction to the left.