Final project: Data analysis and Poster presentation

- Final project information is available on Quercus.
- After today's presentation exercise, take the rest of the tutorial to read about the information, take a look at the dataset you are given and discuss it within your group, start thinking about the final project as soon as possible.
- Forming your group for the final project by next week. Groups will consist of 3-4 students.
- You will present your work at Bahen on Monday, April 1st, so you have less than one month to work on it, again, please get started as early as possible.

Statistical learning / machine learning

- One of the most popular research area in the fields of Statistical science and Computer science, and it has lots of amazing progress and achievement in recent year.
- Three main types:
- Supervised learning. (The only focus for this course).
- Unsupervised learning.
- Reinforcement learning.

- Build a model to make prediction on the future data based on the current data you have (i.e. training data).
- Example: make prediction on the price of house in Toronto based on: the size of house, location, environment, transportation, age, nearby facilities, etc.

- Supervised learning: the data set you have has the label naturally or manually added.
- Some example: linear regression (you will learn in the next few week).; classification tree;
- If you train multiple classification tree on the dataset, and randomly or systematically pick one of prediction among them, then the model you have are called Random forest.

- Unsupervised learning:
- you have only the dataset, and you want to classify them into different categories, i.e. clustering of the data, finding the intrinsic pattern within the data.
- Often useful when the training data don’t have the label (naturally or due to some mistake during your data collection process) and we don’t want to label them manually because that would bring in some human biases.
- Sometime, when both supervised learning and unsupervised learning method are available, unsupervised learning works much better, a good example: classify hand-written digits.

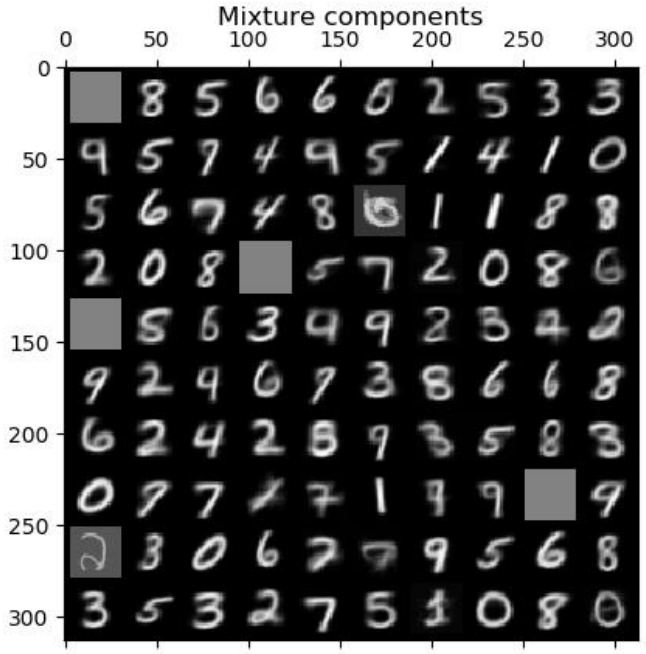
Example: Classify handwritten digits

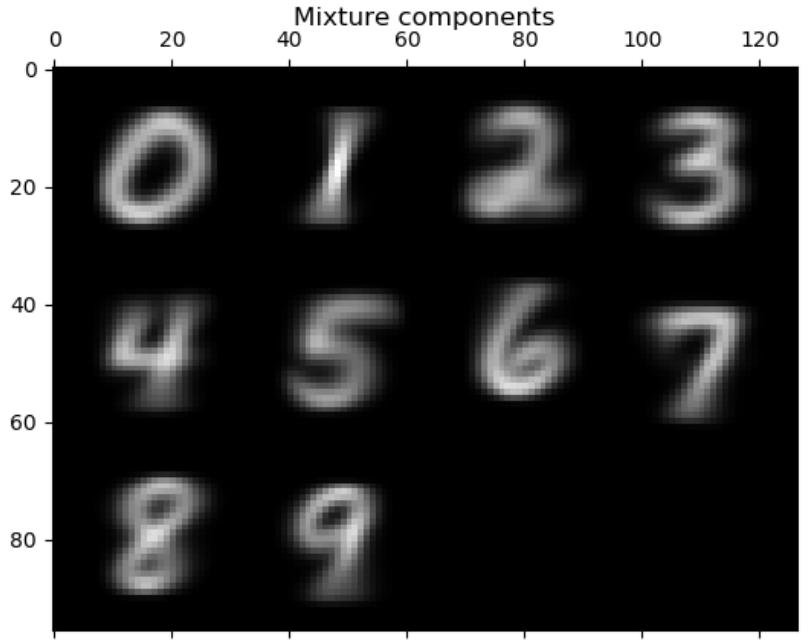
- MNIST dataset of handwritten numbers:

- In the setting of supervised learning we manually label each digit by number in {0,1,2,3,4,5,6,7,8,9}.
- And this is of course the most natural thing to do with the data!
- The task here is to let the computer learn how to distinguish each hand-written digits, i.e. you have a new hand-written digit, tell me what is it.

- Or given only half of the digit, predict the rest and predict which number is it.

- Surprisingly, when we manually assign label to the hand-written digits, the model we learned does not perform better than the case we don't assign label to them! (i.e. in an unsupervised setting) Even if the computer get accessed to more information when assigning labels.
- Some intuition behind it:
- In an unsupervised setting, the computer groups the digits into 100 or more different classes (instead of 10: {0,1,2,…,9} by our manual labeling.)





- So, sometime, by manually labeling data would introduce human biases, or restrict the computation resource that can be used to our learning.
- But, supervised learning also has lots of advantage compared to unsupervised learning.
- E.g. Simper to implement, easier to reason about, faster to train and make prediction.

- Reinforcement learning: You want to train an agent to play a game, you will have a objective function that you want to optimize.

- Training set:

- In the example of predicting trouble sleeping shown on the slides, the **training** refers to determine the threshold of the splitting variables, i.e. if the value of some variable e.g. SleepsHrsNight is lower than some value, then we make a prediction that this person has trouble in sleeping.

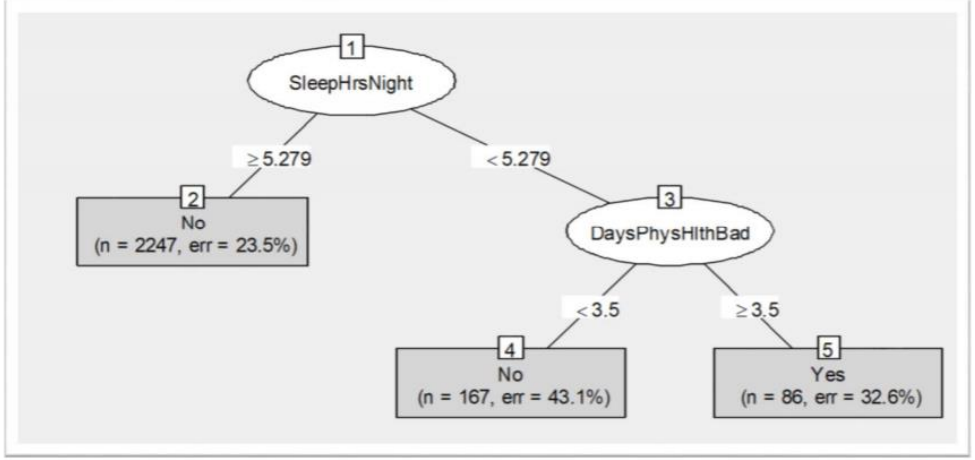
- Testing set: **Only** used it to test the prediction accuracy of your final model, and we NEVER use it at any other stage of the learning.
- Use your model to predict the label of the new data, and compare the prediction to the true label of the data.
- In the example of classification tree, you take the data point and traverse the tree to see which node the point end up with, the label of the node is your prediction.

For those of you who are interested in.

- Besides Training set and testing set, there is one another set called validation set.
- For now, you are given the variables that are used for splitting, but in practice, it is not always going be the case that you know what is the splitting variable.
- E.g. you want to build a classification tree model for a spam filter for email, usually, we don’t know what features (e.g. what words does it contain, i.e. the splitting words) to use to distinguish between a spam email and non-spam email.
- In this case, we need the validation set.

- Validation set: tune hyperparameter, e.g. select the best depth of the classification tree, or select which predictors you want to include in your model (e.g. splitting variables) and prevent overfitting (prevent our model from trying to memorizing the data).
- Here, we are told to use the predictors (variables): SleepHrsNight and DaysPhysHlthBad
- But often in practice, we don't know which predictors or variables should we include in our model, and the job of the Validation set is to help us with that.

```
tree <- rpart(SleepTrouble~SleepHrsNight+DaysPhysHlthBad, data=data)
plot(as.party(tree), gp=gpar(cex=0.8), type="simple")
```



Training

- Trying to figure out at each training iteration What is a "good" split?
- A "good" split is one that makes its child nodes as pure as possible.
- Measuring impurity:

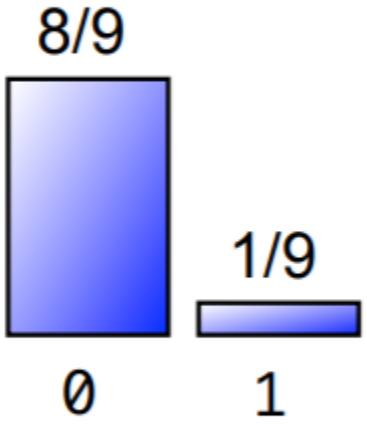
$$Gini(t) = 1 - (w_1(t))^2 + (w_2(t))^2$$

$$Entropy(t) = -w_1(t)\log_2(w_1(t)) + w_2(t)\log_2(w_2(t))$$

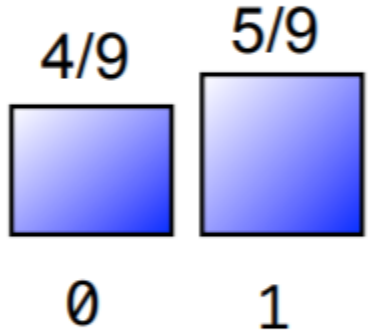
Notes the typo in the slides, here the plus should be minus

Entropy example:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$



$$-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2}$$



$$-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

Some discussion questions

- 1. Can you think of any real-life examples where you may want to develop a classification tree?

- 2. Suppose you were interested in making a classifier to predict what movie somebody would be most interested in. To do this, you first gathered data from a sample of your closest friends. You validated and tested your classifier using different subsets of this data. Now you wish you use your classifier to predict which movie Dr. Moon/ White, your TA, your parents, etc. would like. How well do you think your classifier will perform in these cases?

- 2. *Students' friends will likely be very different from their TA, prof, parents, etc., including different movie preferences and factors that determine their movie preferences.*
- *Take away: you should develop / test your model using datasets that are representative of the population you'd like to apply the classifier to.*
- *Real life example: facial recognition software, came under a lot of scrutiny for validating their algorithms based on largely Caucasian people. This led to many issues with use in the general population, e.g. facial recognition to log-in to your computer. Many big businesses receive a lot of scrutiny in the press/news for this.*

Presentation exercise

- Pick one of the following topics regarding your homework assignment and prepare a **5** minutes presentation.
- Hand in your draft for your presentation.
- As usual, I will assign students to provided feedback to each group.
- Note that, part of your final project is to do review and provide feedback for other group's presentation.

- ***Topic one:***
- Refer to Question 4 from the homework.
- Explain how to make a ROC curve and the type of information it provides.
- Based on the ROC curves provided, describe the accuracy of each of the 3 trees.
- Does this fit your expectations based on the description of how each classifier identified spam emails?

- ***Topic two:***
- Refer to Question 2 from the homework.
- Explain what a confusion matrix is and how each cell is calculated.
- Using the calculated confusion matrix answer the following questions: What percentage of disease positive people who were classified as disease positive were actually disease positive according to cutpoint A? According to cutpoint B?
- What is another term used to describe the percentage you calculated above?

- ***Topic three:***
- Refer to Question 1b.
- Summarize the classification tree from part (b); make sure to include *at least* the following points: how the splits on each variable were selected, how a new observation would be predicted by this classification tree.
- Do you think there may be other important factors to consider? Explain.