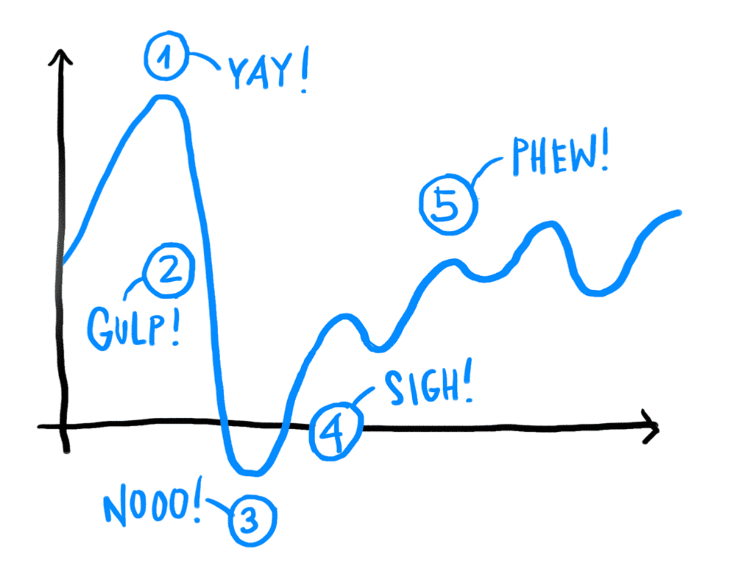
Lecture 24: Recap and Exam Info

COMP90049

Semester 2, 2021

Lida Rashidi and Qiuhong Ke, CIS

A semester of machine learning (?)



Source: `https://www.evalotta.net/blog/2016/4/19/on-practice`

Roadmap

This lecture

- Details on the exam
- Recap of the subject content

Exam Details

Date, time, format...

- The exam will be on **Thursday, November 18th at 3pm**
- The exam will not be invigilated, and it will be an **open book** exam which allows you to use **authorized materials**
- The exam will be **2 hours**, with an additional **15 minutes of reading time**
- The exam will be a **Canvas Assignment**
- The Canvas assignment will be available for an additional 30 minutes after the due time.

Aim to submit on time (!)

Technical difficulties and late submissions

- We accept exams submitted **up to 30 minutes after the due time** with no penalty. This extra time accounts for any technical difficulties during the exam.
- For students with technical difficulty who cannot upload their files via Canvas, you can use this OneDrive link to upload your work.
- Submissions **more than 30 minutes after the due time** can **only be uploaded via the OneDrive link and will result in a deduction of 1 mark from the final mark (not the exam mark) for each minute late up to 30 minutes**. The time stamp on the server will be used as the submission time.

Aim to submit on time (!)

THE UNIVERSITY OF MELBOURNE

Exam Content Details

- Worth **40% of your grade**
- A number of questions of three different categories (coming up next)
- You should attempt all questions (no pick-and-choose)
- Questions have different weight (!)
- **The exam is worth 120 marks**, i.e., $\approx$ 1 mark per minute. The marks associated with a question will give you an idea about how much time you should spend on it.

Exam Format Details

“This is an open book exam. You should enter your answers in a Word document or PDF, which can include typed and/or hand-written answers.

You should answer each question on a separate page, i.e., start a new page for Question 1, Question 2, etc parts within questions do not need new pages. Write the question number clearly at the top of each page.

You have unlimited attempts to submit your answer-file, but only your last submission is used for marking. If your last submission arrives more than 30 minutes after the due time, you will lose 1 mark from the final mark (not the exam mark) for each minute late up to 30 minutes.”

Exam Format Details

Authorised materials:	Lecture slides, workshop materials, prescribed reading, your own project reports.
Calculators:	Permitted

You must not use materials other than those authorised above. You are not permitted to communicate with others for the duration of the exam, other than to ask questions of the teaching staff via the Big Blue Button tool in Canvas. Your computer, phone and/or tablet should only be used to access the authorised materials, enter or photograph your answers, and upload these files. The work you submit **must be based on your own knowledge and skills**, without assistance from any person or unauthorized materials.

Section A: Short answer Questions

- Requiring you to explain or compare concepts covered in this subject.
- some may require a small amount of calculation
- to be answered in 1-3 (handwritten) lines, unless otherwise instructed

Section A: Short answer Questions

- Requiring you to explain or compare concepts covered in this subject.
- some may require a small amount of calculation
- to be answered in 1-3 (handwritten) lines, unless otherwise instructed

Section A: Short answer Questions [40 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each question in 1-3 lines, with longer responses expected for the questions with higher marks.

Question 1: [40 marks]

(a) Name three differences between exact optimization and Gradient descent. [6 marks]

(b) Indicate the best alignment of the concepts under (a) to the concepts under (b). Many-to-one and one-to-many alignments are possible. [3 marks]

(a)	(b)
clustering	supervised
classification	semi-supervised
regression	unsupervised

Section B: Method Questions

- **Resembling Workshop Questions**
- demonstrate your conceptual understanding of the methods that we have studied in this subject.
- usually involve some calculations, and you will need to show your calculations (i.e., not just state the answer)

Section B: Method Questions

- **Resembling Workshop Questions**
- demonstrate your conceptual understanding of the methods that we have studied in this subject.
- usually involve some calculations, and you will need to show your calculations (i.e., not just state the answer)

Section B: Method & Calculation Questions [55 marks]

In this section you are asked to demonstrate your conceptual understanding of methods that we have studied in this subject, and your ability to perform numeric and mathematical calculations.

Question 2: K-Nearest Neighbors [8 marks]

With respect to the following data set of 6 instances with 3 attributes and two classes F and T, plus a single test instance labelled "?":

instance #	ele	fed	aus	CLASS
1	1	1	1	F
2	1	0	0	F
3	1	1	0	T
4	1	1	0	T
5	1	1	1	T
6	1	1	1	T
7	0	0	0	?

Explain why a model with $K = 1$ will make a different prediction compared to a model with $K = 3$ on the given test instance. You do not need to show your work for this question, but should provide an explanation which refers to the data.

Section C: Design and Application Questions

- **Resembling Assignment Questions**
- demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding.
- Expected answer to each question to be from one third of a page to one full page in length (hand-written).
- Require significantly more thought than Sections A or B, and should be attempted last.

Section C: Design and Application Questions

Question 10: Insurance Policy [25 marks]

You are a manager of a life insurance company and want to provide optimal insurance quotes to your potential customers. The quotes fall into one of three categories 'high', 'medium' or 'low' premium. Your company is so popular that you cannot sort through all applications manually. Instead, you wantto pre-sort applications into meaningful groups. Each application comes with features such as

- Name of applicant
- Age of applicant
- Favorite color of applicant
- Longest period spent in hospital
- Marital status of applicant
- Gender of applicant

Please answer the following questions with respect to the machine learning problem introduced above.

1. Describe the machine learning concept and features underlying this task. [3 marks]
2. Assume you have access to the following ML methods: (a) Decision trees; (b) neural networks; (c) k-means. For each algorithm, state whether it is appropriate in this situation as well as a reason for your decision [6 marks]
3. Now assume a slightly different situation where you (a) have access to a set of 50 admission decisions from previous years. Describe how this new information will change (a) your machine learning approach. [8 marks]
4. Further questions e.g., on evaluation or feature selection ... [3 marks]

Recap part I: Basic Concepts in Machine Learning

"We are drowning in information, but we are starved for knowledge"

John Naisbitt, Megatrends

Our definition of Machine Learning

automatic extraction of **valid, novel, useful and comprehensible knowledge** (rules, regularities, patterns, constraints, models, ...) from arbitrary sets of data

Three ingredients for machine learning

Data

- Discrete vs continuous vs ...
- Big data vs small data
- Labeled data vs unlabeled data
- Public vs sensitive data

Models

- function mapping from inputs to outputs
- parameters of the function are unknown
- probabilistic vs geometric models

Learning

- Improving (on a task) after data is taken into account
- Finding the best model parameters (for a given task)
- Supervised vs. unsupervised

Terminology

- The input to a machine learning system consists of:
  - **Instances**: the individual, independent examples of a concept, also known as **exemplars**
  - **Attributes**: measuring aspects of an instance also known as **features**
  - **Concepts**: things that we aim to learn generally in the form of **labels** or **classes**

Instance Topology

- Instances characterised as “feature vectors”, defined by a predetermined set of attributes

- Input to learning scheme: set of instances/dataset
  - Flat file representation
  - No relationships between objects
  - No explicit relationship between attributes

Instance Topology

- Instances characterised as “feature vectors”, defined by a predetermined set of attributes
- Input to learning scheme: set of instances/dataset
  - Flat file representation
  - No relationships between objects
  - No explicit relationship between attributes
- Possible attribute types (levels of measurement):
  1. nominal
  2. ordinal
  3. continuous

Also: Feature Selection Why? How?

THE UNIVERSITY OF MELBOURNE

Recap part II: Classification

Classification

- **Lazy learning:**
  No model training, Instance-based Learning
- **Eager learning:**
  Train a model using training data and use the model to predict test instances

K-NN

- K-NN is a lazy learner

- Algorithm
  - Measure the similarity (or distance) between the test instance and training data
  - Find K-Nearest neighbours
  - Return the class of the test instance using the corresponding labels of the K-Nearest neighbours
- Advantages:
  - Intuitive
  - No assumptions
  - Evolve and adapt immediately
- Concerns:
  - What similarity (or distance) measure?
  - How to aggregate the labels of the neighbours?
  - What K value?
  - Expensive if the data set is large

Eager learning

- **Linear Classification**
- **Non-Linear Classification**

Recap part II: Linear Classification

Naive Bayes I

Task: classify an instance $D = \langle x_1, x_2, ..., x_n \rangle$ according to one of the classes $c_j \in C$

$$
\begin{aligned}
c &= \underset{c_j \in C}{\operatorname{argmax}} P(c_j | x_1, x_2, ..., x_n) && (1)\\
&= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(c_j) P(x_1, x_2, ..., x_n | c_j)}{P(x_1, x_2, ..., x_n)} && (2)\\
&= \underset{c_j \in C}{\operatorname{argmax}} P(c_j) P(x_1, x_2, ..., x_n | c_j) && (3)\\
&= \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_i P(x_i | c_j) && (4)
\end{aligned}
$$

Posterior $P(c_j | x_1, x_2, ..., x_n) = \frac{prior * likelihood}{evidence}$

What does the equality between (3) and (4) imply?

The problem with unseen features

- If any term $P(x_m|y) = 0$ then the class probability $P(y|x) = 0$
- **Solution:** no event is impossible: $P(x_m|y) > 0 \forall x_m \forall y$
  1. Epsilon Smoothing
  2. Laplace Smoothing

The problem with unseen features

- If any term $P(x_m|y) = 0$ then the class probability $P(y|x) = 0$
- **Solution:** no event is impossible: $P(x_m|y) > 0 \forall x_m \forall y$
  1. Epsilon Smoothing
  2. Laplace Smoothing

Estimation

Question 3: Naive Bayes [5 marks]

Name the optimization strategy you would choose to estimate the parameters of a Naive Bayes model. Compare the strategy against an alternative strategy, and provide two reasons why your chosen strategy is preferred.

Logistic Regression

- Is a **binary** classification model
- Is a **probabilistic discriminative model**. Why?
- We model **probabilities** $P(y = 1|x; \theta) = p(x)$ as a function of observations x under parameters θ. [What about $P(y = 0|x; \theta)$?]
- We want to use a (suitably modified) **regression** approach

Logistic Regression

- Is a **binary** classification model
- Is a **probabilistic discriminative model**. Why?
- We model **probabilities** $P(y = 1|x; \theta) = p(x)$ as a function of observations x under parameters θ. [What about $P(y = 0|x; \theta)$?]
- We want to use a (suitably modified) **regression** approach

(f) Consider the following two tasks: (1) predicting whether a job applicant is successful based on the characteristics of their CV; (2) Predicting the expected salary of a job applicant based on the characteristics of their CV. (i) For each task, (i) name the corresponding machine learning concept. (ii) Justify your choice. [3 marks]

Logistic Regression

- Is a **binary** classification model
- Is a **probabilistic discriminative model**. Why?
- We model **probabilities** $P(y = 1|x; \theta) = p(x)$ as a function of observations x under parameters θ. [What about $P(y = 0|x; \theta)$?]
- We want to use a (suitably modified) **regression** approach

$$P(y = 1|x_1, x_2, ..., x_F; \theta) = \frac{1}{1 + \exp(-(\theta_0 + \sum_{f=1}^{F} \theta_f x_f))} = \sigma(x; \theta)$$

- We define a **decision boundary**, e.g., predict $y = 1$ if $P(y = 1|x_1, x_2, ..., x_F; \theta) > 0.5$ and $y = 0$ otherwise

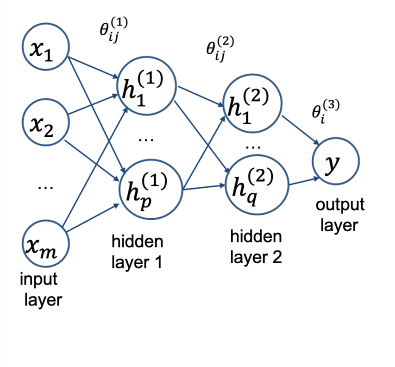
Perceptron: Definition I

- The Perceptron is a **minimal neural network**
- **Neural networks** are inspired by the brain – a complex net of **neurons**
- A (computational) neuron is defined as follows:
  - input = a vector x of numeric inputs ($\langle 1, x_1, x_2, ...x_n \rangle$)
  - output = a scalar $y_i \in \mathbb{R}$
  - hyper-parameter: an **activation function** f
  - parameters: $\theta = \langle \theta_0, \theta_1, \theta_2, ...\theta_n \rangle$
- Mathematically:

$$y^i = f\left(\left[\sum_j \theta_j x_j^i\right]\right) = f(\theta^T x^i)$$

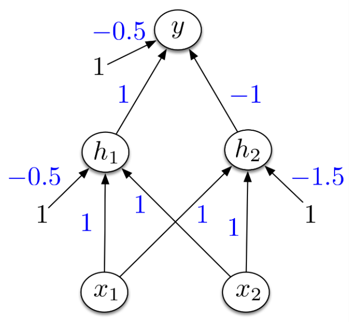
Recap part III: Non-Linear Classification

- **Input layer** with input units x: the first layer, takes features/raw data x as inputs
- **Output layer** with output units y: the last layer, has one or multiple units (e.g., 1 unit for binary classification)
- **Hidden layers** with hidden units h: all layers in between.



Multi-layer Perceptron: Prediction

What is the output given an input pair, e.g., $x1 = 1$ and $x2 = 0$?

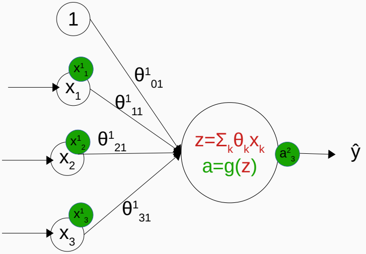


$\phi(x) = \begin{cases} 1 & \text{if } x >= 0 \\ 0 & \text{if } x < 0 \end{cases}$ and recall: $h_j^{(l)} = \phi^{(l)}\left(\sum_i \theta_{ij}^{(l)} h_i^{(l-1)} + \theta_j^{(l)}\right)$

Source: `https://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/readings/L05%20Multilayer%20Perceptrons.pdf`

Recall Perceptron learning:

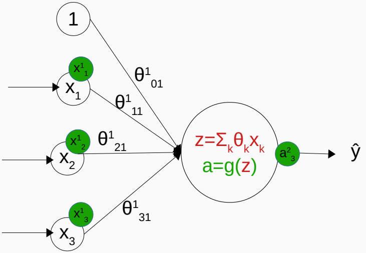
- Pass an input through and compute $\hat{y}$
- Compare $\hat{y}$ against y
- Weight update $\theta_i \leftarrow \theta_i + \eta(y - \hat{y})x_i$



Learning the Multi-layer Perceptron

Recall Perceptron learning:

- Pass an input through and compute $\hat{y}$
- Compare $\hat{y}$ against y
- Weight update $\theta_i \leftarrow \theta_i + \eta(y - \hat{y})x_i$

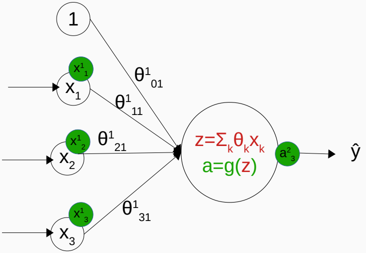


Why can't we use this method to learn parameters of the MLP?

What do we do instead?

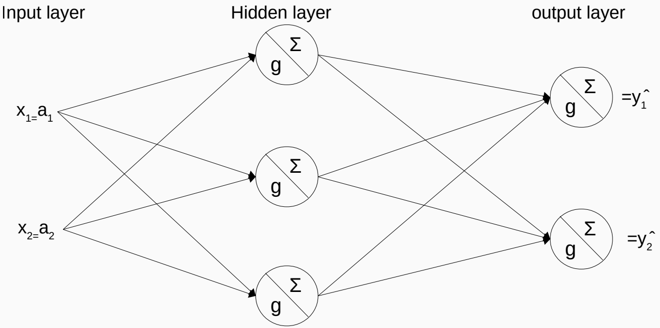
Recall Perceptron learning:

- Pass an input through and compute $\hat{y}$
- Compare $\hat{y}$ against y
- Weight update $\theta_i \leftarrow \theta_i + \eta(y - \hat{y})x_i$

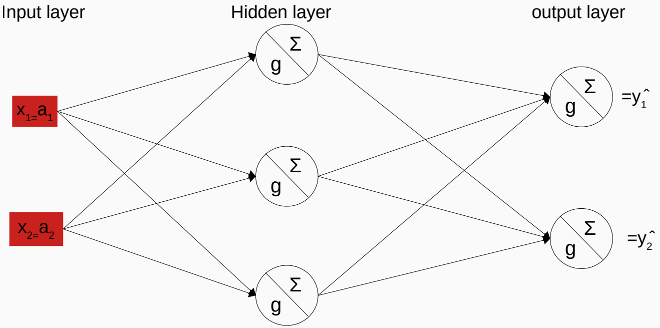


- The above δ_i can only be applied to output units, because it relies on the **target outputs** y^p.
- We do not have target outputs for the intermediate hidden layers in the MLP
- Use Backpropagation

Backpropagation: Demo

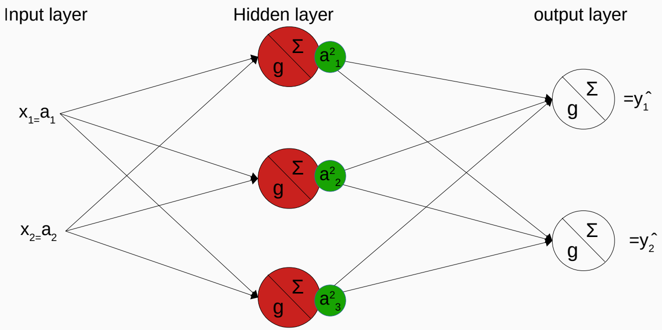


Backpropagation: Demo



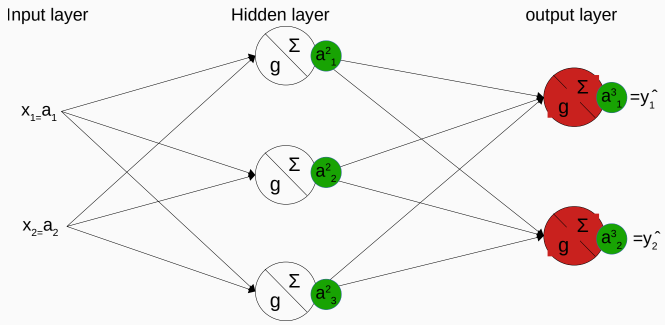
- Receive input

Backpropagation: Demo



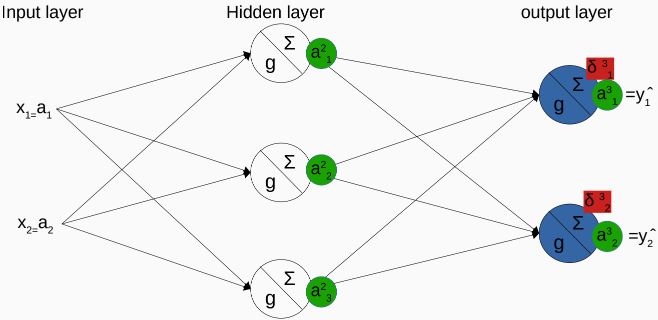
- Receive input
- Forward pass: propagate activations through the network

Backpropagation: Demo



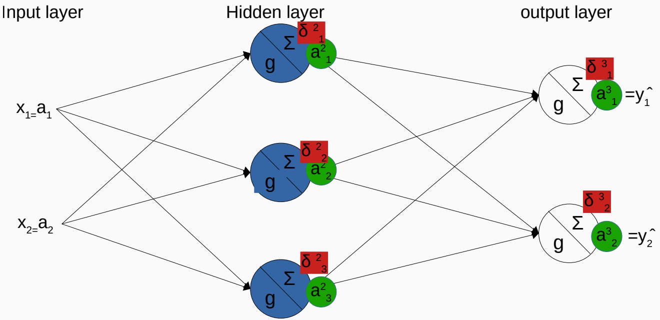
- Receive input
- Forward pass: propagate activations through the network

Backpropagation: Demo



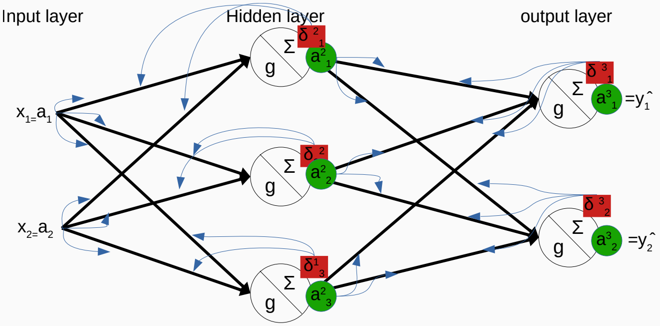
- Receive input
- Forward pass: propagate activations through the network
- Compute Error of output layer: $\delta_o^{out} = \frac{\partial L}{\partial a_o^{(out)}} g'$

Backpropagation: Demo



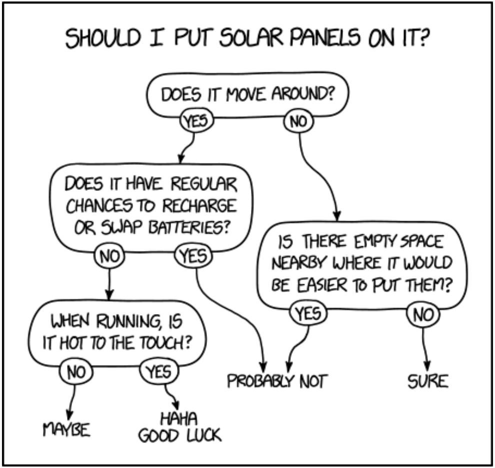
- Receive input
- Forward pass: propagate activations through the network
- Compute Error of output layer: $\delta_o^{out} = \frac{\partial L}{\partial a_o^{(out)}} g'$
- Backward pass: propagate error terms through the network
  $\delta_j^{(l)} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} \theta_{jk}^{(l+1)} g'$

Backpropagation: Demo



- Receive input
- Forward pass: propagate activations through the network
- Compute Error of output layer: $\delta_o^{out} = \frac{\partial L}{\partial a_o^{(out)}} g'$
- Backward pass: propagate error terms through the network
  $\delta_j^{(l)} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} \theta_{jk}^{(l+1)} g'$
- Calculate $\frac{\partial L}{\partial \theta_{ij}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)}$ for all $\theta_{ij}^{(l)}$
- Update weights $\theta_{ij}^{(l)} \leftarrow \theta_{ij}^{(l)} - \eta \frac{\partial L}{\partial \theta_{ij}^{(l)}}$

Decision Trees



https://xkcd.com/1924/

- ID3 **algrithm** to construct decision trees
- Split **criteria**:
  - entropy
  - information gain
  - gain ratio

Ensemble learning (aka. Classifier combination):

Methods

- Stacking
- Bagging (Random Forests)
- Boosting (Adaboost)

Recap part IV: More Food for Thought (or exam preparation...)

Questions to think about I

Choosing a classification (or any ML) Algorithm

- Probabilistic interpretation?
- Restrictive assumptions on features?
- Restrictive assumptions on the problem?
- How well does it perform?
- How long does it take to train?
- How interpretable is it?
- How much data does it require?

Questions to think about II

How do we know we succeeded?

- Choose the right evaluation metric (accuracy, precision, recall, ...)
- Know the mechanics behind the metrics.
- What is **overfitting** and how do we prevent it?
- Choose the right evaluation strategy, maximizing the utility of your data (cross-validation, hold-out, ...). What to consider?

How do we know we succeeded?

(d) [3 marks] Consider the following set of evaluation metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{Error Rate} = 1 - \text{Accuracy}$$

1. What types of machine learning algorithms can be evaluated with these measures? [1 mark]
2. Explain why. [2 marks]

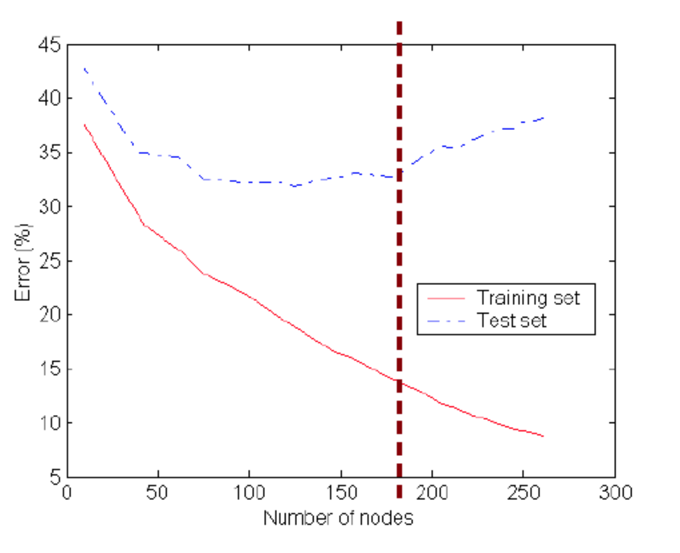
Questions to think about III

Theoretical considerations and optimization

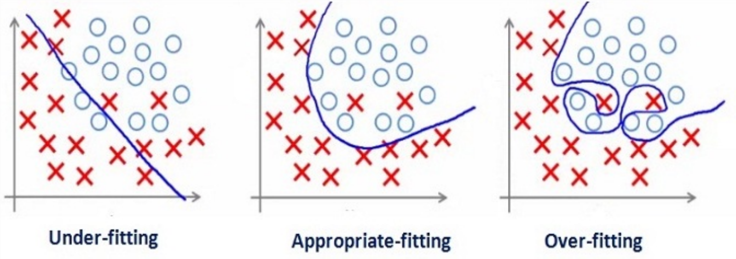
- Is the problem linearly separable?
- Is my classifier powerful enough to solve my problem?
- What does the objective function of my classifier look like? And what optimization strategy should I choose?

Recap part V: Evaluation

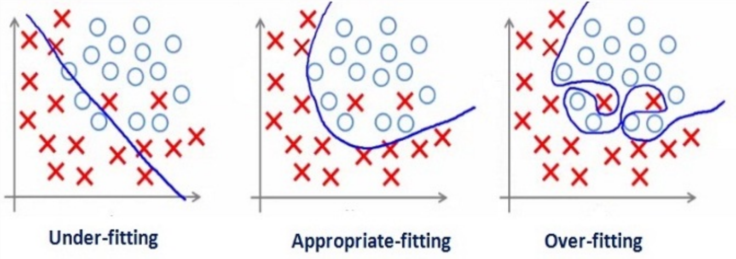
Learning Curves



Underfitting and Overfitting



Underfitting and Overfitting



High Bias

- Use more complex model (e.g. use nonlinear models)
- Add features
- Boosting

High Variance

- Reduce model complexity – complex models are prone to high variance
- Reduce features ; add data
- Bagging

Recap part VI: Beyond supervised learning...

Active learning/Semi-supervised learning

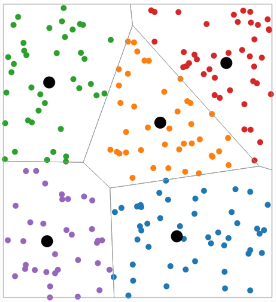
- Active learning: query strategies
- Semi-supervised learning: Self-training
- Active learning vs Semi-supervised learning

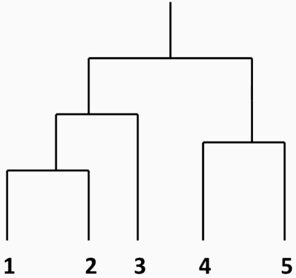
Unsupervised Learning: Clustering

Learning in the context where we *don't* have (or don't use) training data labelled with a class value for each instance.

Finding groups of items that are *similar*.

- *k*-means clustering
- hierarchical clustering
    - agglomerative clustering
    - divisive clustering





Anomaly Detection

Types of Anomalies

- Global, contextual, collective anomalies

Concepts/scenarios of anomaly detection

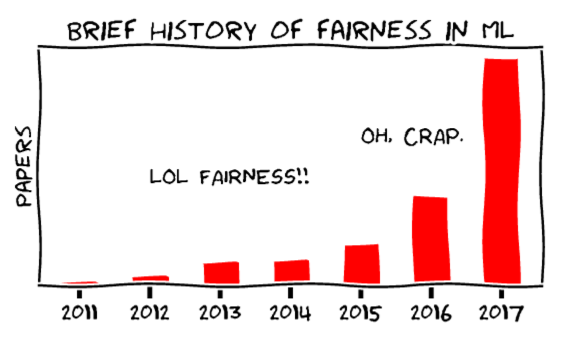
- unsupervised, semi-supervised, supervised methods

Methods

- Statistical methods: assume data follow a fixed model
- Proximity based: outlier if nearest neighbors are far away
- Density based: outlier, if in region of low density
- Clustering based: outlier, if not part of large and dense cluster

Recap part VII: Problems and applications, more generally...

Fair Machine Learning



https://fairmlclass.github.io/

Sources of bias

- Data
- Users
- Models and algorithms

Algorithmic Fairness

- Fairness through unawareness (**Why (not)?**)
- Fairness through awareness: group fairness, equal opportunity, predictive parity

Approaches towards preventing bias in ML models

- Pre-processing, **for example, ...**
- Modeling, e.g., **for example, ...**
- Post-processing, e.g., **for example, ...**

THE UNIVERSITY OF MELBOURNE

Summary

Source https://www.aitrends.com/machine-learning/here-are-six-machine-learning-success-stories/

Summary

Source https://www.aitrends.com/machine-learning/here-are-six-machine-learning-success-stories/

- Understand fundamental mathematical concepts in machine learning (including probability and optimization)
- Understand the theory behind a variety machine learning algorithms
- Identify the correct ML model given a specific data set
- Meaningfully evaluate the output of a ML model in the context of a specific problem
- Apply a variety of ML algorithms
- Python programming: ML model implementation, data processing, evaluation
- Problem solving, Academic writing and presentation

And finally...

Please participate in the university feedback survey!

- What worked well?
- Suggestions for improvements?

Capstone / PhDs

Lida: I am looking for motivated master (capstone) and PhD students, interested in working on Graph-theoretical as well as Information Retrieval Problems. Feel free to get in touch if you're interested!