Lecture 24: Recap and Exam Info

COMP90049
Semester 1, 2024

Ting Dang, CIS

Acknowledgement: Lea Frermann

Roadmap

This lecture

- Details on the exam
- Recap of the subject content

Exam Details

Time, format...

- Duration: **2 hours**, with an additional **15 minutes of reading time**
- Format: **Closed-book, pen-and-paper** exam
- You are allowed to use **non-programmable calculators.**
- The formula associated with each question will be provided.
- The only formulas to memorise are **Accuracy, Precision, Recall, F1-score and entropy**.

Exam Content Details

- Worth **50% 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 100 marks**.

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) Align the concepts under (a) to their most typical type of supervision under (b). [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, or (less commonly) describe the logical process with which you arrived at an answer (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, or (less commonly) describe the logical process with which

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

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

What is 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

- Possible attribute types (levels of measurement):
  1. nominal
  2. ordinal
  3. continuous

Also: Feature Selection Why? How?

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 $c_j \in C$

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

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

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** $P(y = 1|x; \theta)$ 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

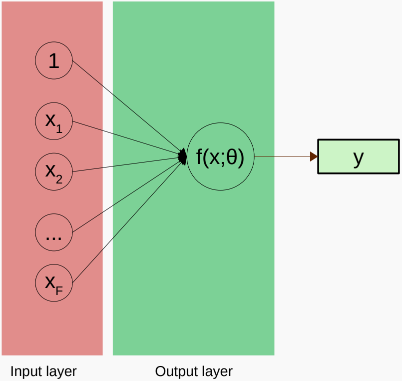
- The Perceptron is a **minimal neural network**
- **Neural networks** are inspired by the brain – a complex net of **neurons**
- A (computational) neuron is defined as follows:
  - input = a vector x of numeric inputs ($\langle 1, x_1, x_2, ... x_n \rangle$)
  - output = a scalar $y_i \in \mathbb{R}$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

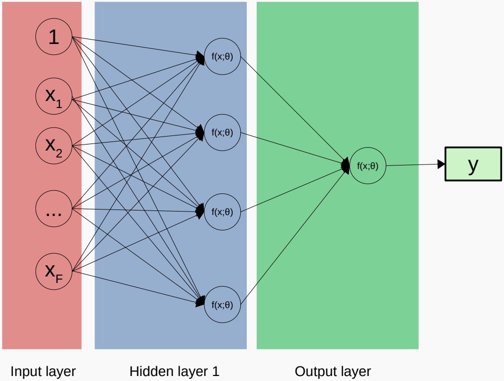
Multi-layer Perceptron I

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



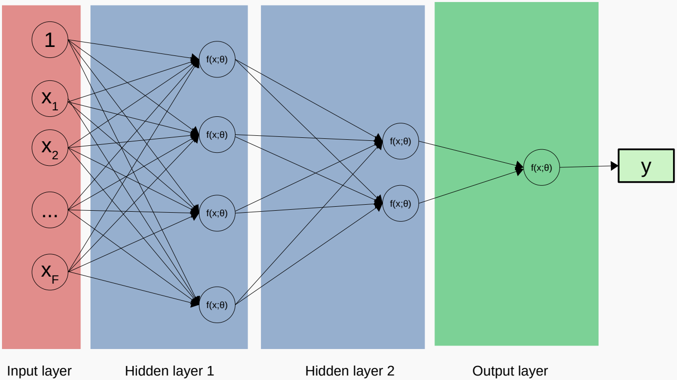
Multi-layer Perceptron I

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



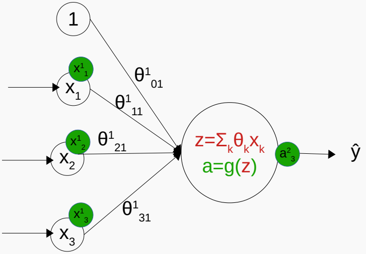
Multi-layer Perceptron I

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



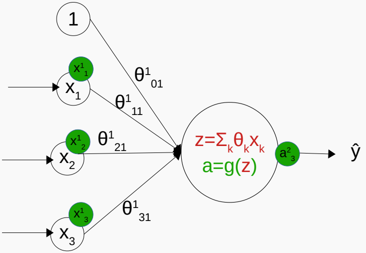
Recall Perceptron learning:

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



Recall Perceptron learning:

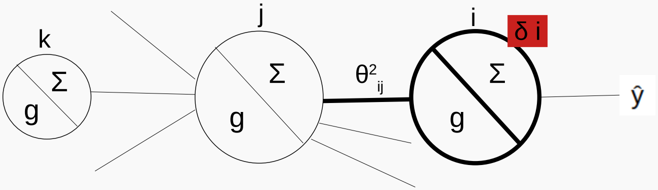
- Pass an input through and compute $\hat{y}$
- Compare $\hat{y}$ against y
- Weight update $\theta_i \leftarrow \theta_i + \eta(y - \hat{y})x_i$



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

What do we do instead?

Backpropagation: The Generalized Delta Rule



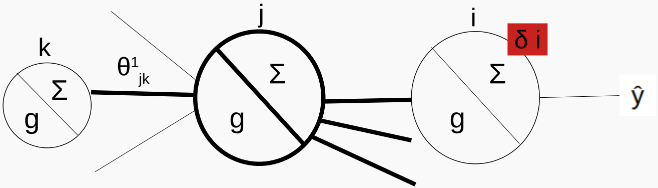
- The Generalized Delta Rule

$$\triangle\theta_{ij}^{2} = \eta\frac{\partial E}{\partial\theta_{ij}^{2}} = \eta(y^{p} - \hat{y^{p}})g'(z_i)a_j = \eta\,\delta_i\,a_j$$

$$\delta_i = (y^{p} - \hat{y^{p}})g'(z_i)$$

- 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 y for the intermediate layers

Backpropagation: The Generalized Delta Rule

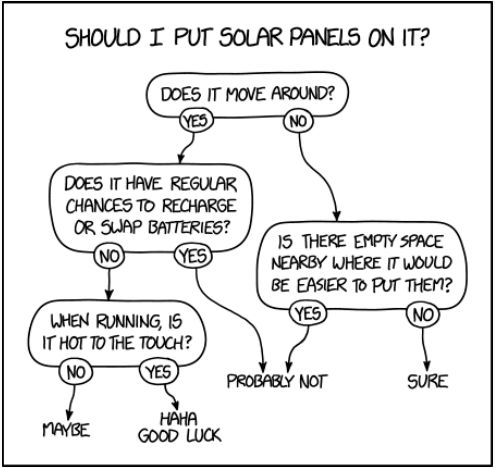


- Instead, we **backpropagate** the errors (δs) from right to left through the network

$$\triangle\theta^1_{jk} = \eta\ \delta_j\ a_k$$

$$\delta_j = \sum_i \theta^1_{ij}\ \delta_i\ g'(z_j)$$

Decision Trees



https://xkcd.com/1924/

- ID3 **algrithm**: recursive divide and conquer
- Split **criteria**:
  - entropy/purity: **intuition?** What's a good value of entropy?
  - information gain
  - gain ratio

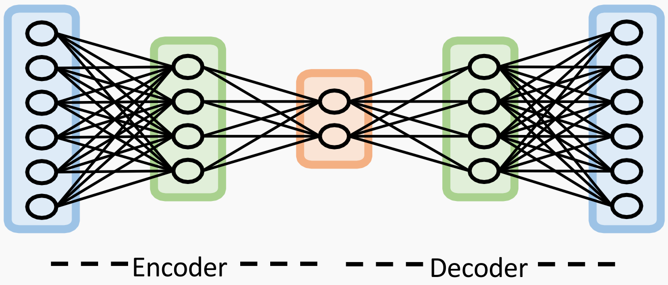
Ensemble learning (aka. Classifier combination): constructs a set of base classifiers from a given set of training data and aggregates the outputs into a single meta-classifier

- **Intuition 1**: the combination of lots of weak classifiers can be at least as good as one strong classifier
- **Intuition 2**: the combination of a selection of strong classifiers is (usually) at least as good as the best of the base classifiers

Methods

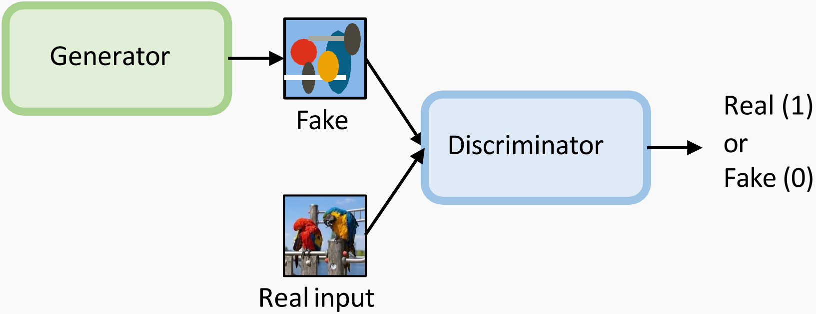
- Stacking
- Bagging (Random Forests)
- Boosting (Decision Trees, Adaboost)

Autoencoders



- Output is the input
- “Bottleneck” layer – smaller than the input, to learn important and representative information
- Unsupervised learning

Generative Adversarial Networks (GANs)



- Learning a function to map the samples from a simple distribution to the distribution of the target
- Training can be difficult – hard to find a balance between discriminator and generator
- Difficult to evaluate: realism, memorization, diversity.

THE UNIVERSITY OF MELBOURNE

Recap part IV: Practical considerations

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]

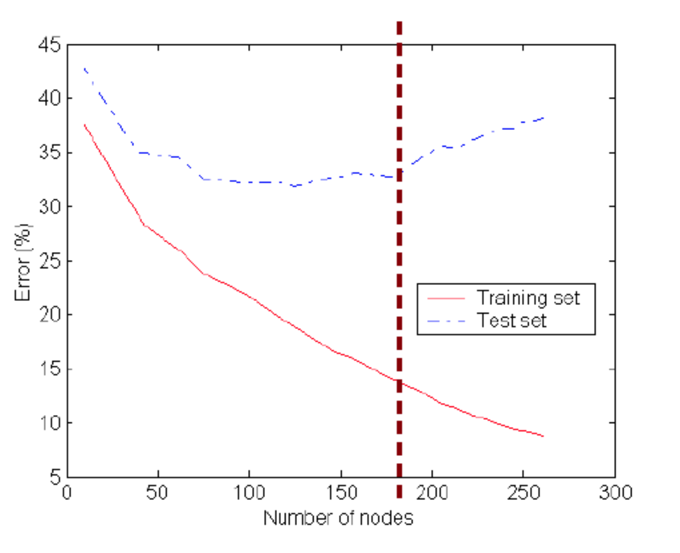
Questions to think about III

Theoretical considerations and optimization

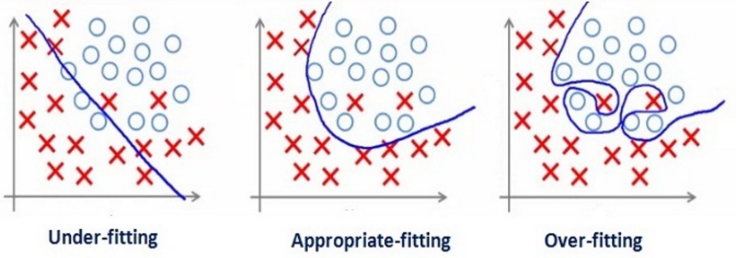
- Is the problem linearly separable?
- Is my classifier powerful enough to solve my problem?
- What does the objective function of my classifier look like? And what optimization strategy should I choose?

Recap part V: Evaluation

Learning Curves



Underfitting and Overfitting



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...

Semi-supervised learning

Learning from both labelled and unlabeled data

- Semi-supervised classification:
  - L is the set of labelled training instances $\{x_i, y_i\}_{i=1}^{l}$
  - U is the set of unlabeled training instances $\{x_i\}_{i=l+1}^{l+u}$
  - Often $u \gg l$
  - Goal: learn a better classifier from $L \cup U$ than is possible from L alone

Approaches

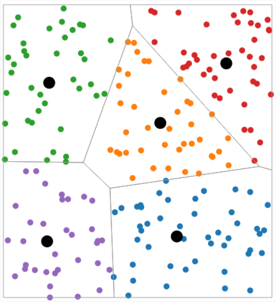
- Self-training
- Active learning, query strategies
- Data augmentation
- Unsupervised pre-training

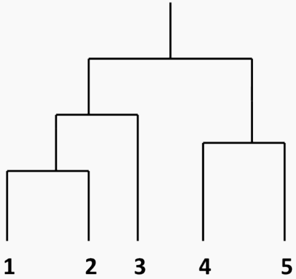
Unsupervised Learning: Clustering

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





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

Anomaly Detection

Types of Anomalies

- Global, contextual, collective anomalies

Concepts/scenarios of anomaly detection

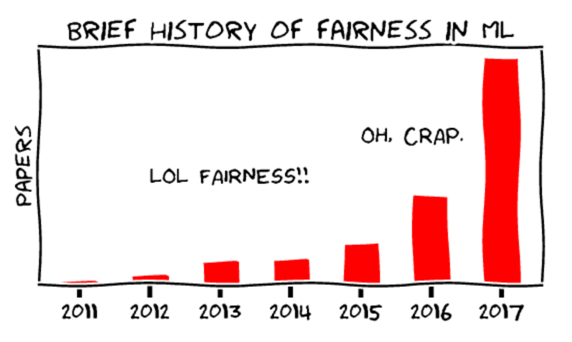
- unsupervised, semi-supervised, supervised methods

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

Name a statistical and a proximity-based method

Fair Machine Learning



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

- 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

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

And finally...

Please participate in the end of semester survey!

- What worked well?
- Suggestions for improvements?

Consider taking subjects COMP90042 (Natural Language Processing) and COMP90051 (Statistical Machine Learning) to build on the skills you gained this semester!

All the best!