Every time the professor mentioned something about the exam

Silvanus Bordignon

- If you want the full notes for the lectures in Italian look at the file Appunti per Introduction to Machine Learning 2023-2024, if you only want the questions and answers for the oral exam in Italian look at the file domandeML.
- NOT everything you need to know to pass the exam, BUT everything she explicitly mentioned she usually asks at the exam
- sorted by lecture/pdf in which she mentioned it
- kinda helpful (?), looks like she wants basically everything except the most complicated maths

Introduction

ML basics (Lecture 2)

- we won't be covering much of data preprocessing, meaning that we need to know what it is, but that's pretty much it
- dimensionality reduction is one of her favourite questions
- for each algorithm we'll see we need to understand the pros and cons; when we have a problem to solve, which one should we use and why
- describe what is a learning process (training phase, test phase, how they're connected, training phase describe the challenges; made of several blocks, sometimes we need to process data, pipeline, and so on)
- what is the mapping from data to features and role of models into this data
- difference between supervised, unsupervised and reinforcement learning
- what does generalization mean? What does data generating distribution mean, and why is it important

ML basics (Lecture 3)

- we need to understand the challenges that arise at each step of the ML pipeline: task, data, model, objective, algorithm; each step has its own complexities
- what is the hypothesis space? Give an example
- overfitting, underfitting and fighting generalization
- examples of overfitting and underfitting with regression, classification, etc.
- we DO NOT reduce the training data to prevent overfitting. We apply techniques like regularization.

KNN

- slide 16 would be how you would start answering the question: "What's the KNN algorithm"

- KNN is naively designed to handle the multi-class classification case, while some other algorithms need to be adapted
- pros and cons of KNN

Linear models

Multi-class classification

- on slide 38 is how she wants the OVA algorithm explained; she said that the intuition is important, but we should also be able to express formally how this approach works
- in the context of slide 48/49, it's preferrable to learn more models on smaller data points that few models on large training sets; this is especially true for SVMs, since each training is finding a solution for quadratic optimization problem, which can take quite some time
- difference between microaveraging and macroaveraging

Gradient descent

- what's a hyperplane? A hyperplane is a function that takes the features and multiples then by the corresponding weights, adding a bias
- what is the relationship between the perceptron learning algorithm and gradient descent? We apply very similar updates on the weights, depending on labels and features; difference, the c value we have discussed, and how often theese updates are performed

SVM

- what is the formulation of the optimization problem for linearly separable SVMs? Substituting the question of the margin, and imposing that all the points are correctly classified; maximizing the margin subject to a set of constraints.
- (after the first part, the linearly-separable case) no need to know all formulas, just maybe equation hard margin SVM, equation soft margin, what they mean. Each part is important in the equation, but no need to know ALL formulas
- two main innovations: the notion of margin and the idea of the kernel trick, making them able to address linear and non-linear problems in the same framework

DT

- what is a decision tree? The description can be found on slide 6
- no need to remember all the formulas for decision trees (different impurities and so on); undestand the flavor of these slides, what they mean, why a method is better than the other one and so on
- loves asking pros and cons of the algorithms and models we study

Unsupervised learning

- know the definition of the three main tasks we saw in unsupervised learning: dimensionality reduction, clustering, density estimation
- slide 13, all the Maths is not asked at the exam
- about principal component analysis
    - in PCA, how do I compute the principal component?
    - in PCA, how do I choose the number of principal components?

    - what's the main limination of PCA? It's a **linear** technique for dimensionality reduction

Clustering

- illustrate the k-means algorithm
- how do you handle non-gaussian data?

Introduction to Neural Networks

- backpropagation, you can start with: it's a procedure to compute the gradient needed to train the neural networks and is composed of three steps: forward propagation, error estimation and the backpropagation itself
- neural networks tackle complex non-linear surfaces by appending many many layers, SVMs use the kernel trick, which works very well; that's one of the main reasons why SVMs were the method of choice
    - also because, given a god feature extractor, they are very powerful objects
    - and they are far easier to train, with way less hyperparameters
    - and they give you a theoretical generalization guarantee, something we do not have in NN
- no need to memorize all the backpropagation formulas, undestand the flow; need to know what is backpropagation and why it works
- explain the three steps of backpropagation and why they work

Neural Networks II

- difference bwteen BDF and SGD; answer on slide 15 of the second pdf on NN
- momentum in SGD
- for GoogleNet, and all the single architectures in this slide deck, she only wants what she said in the lecture, no need no need to memorize each architecture in detail
- remember than feed-forward neural networks, CNNs, are objects that we first saw for supervised learning, meaning we need images and their annotations
- differences between AE and VAE, main one being that AE are used for dimensionality reduction while VAE for density estimation and sample generation

Generative models

- are classic autoencoders already generative models? No, by default, the decoder taken from a classical autoencoder is not able to generate reasonable data
- understand the main idea of VAE, of how we try to reach a certain subset of the distribution, the variational bound part and which is the objective function I'm using to train VAE
- the maths on slide 15 can be replaced by: "for our objective function we want to minimize the KL-divergence between these two data distributions; since it cannot be solved exactly, to approximate it we use an upper bound made of two terms: one of them is a reconstruction term, similar to an autoencoder in a probabilistic fashion, making sure we cover the space of our data; second term, KL divergence between the two distributions that can be computed in closed form, and making sure our latent space is smooth, so that when moving in this space we can generate data that sits "inbetween" (i.e. grinning face between neutral and smile)."
- slide 22
- in the advantages of VAE include always the smooth latent space and the speed in generating new samples

- slide 31, won't ask the specific formula but know how we get the objective function
- of all the maths, what's important to know for the exam is slide 36 (and 37)
- GANs are also fast at inference time
- why should we use a GAN and why should we use a VAE

Diffusion models

- slide 8, how does a generative model generate new data?
- what she cares about is that we understand the forward/backward process on slide 30, the algorithms on slide 34 and the structure of the network on slide 35

Reinforcement learning

- what is RL
- give me an example of an application that requires RL
- what is a Markow Decision process and why I need it (I need the MDP loop in order to compute the process)
- what is cumulative discounted reward
- supervised vs RL (default, and in NN)
- value function encoding and why the Bellman equation is important (allows me to build my Q-value function)
- why at a certain point we need a NN (scale, handle problems also requiring huge action spaces)
- intuition of policy gradient method

Q&A session

- why are VAE an explicit density estimation approach?
- how does a kernel function work, mapping to a higher dimension?
- what's the kernel for, provide an example of a kernel, formulation of the dual SVMs, so where the kernel intervenes
- polynomial regression, what's the impact of the degree of the polynomial on the fitting? What was the hypothesis space in that example?
- difference between parameter and hyperparameter
- is the Hinge loss an object for classification or regression?
- where do we talk about the notion of margin? (linear models!)
- KNN, how do I regulate overfitting? What is a Voronoy partition? Notion of "nearest"?
- which are the problems of the Euclidian distance?
- curse of dimensionality?
- what does it mean doing gradient descent?
- what's the learning rate? Impact? Why do we need to change it? What happens if it's too big or too small?
- effect of applying BGD or SGD? (smooth path progressing to the minimum vs oscillations; mini-batches also solve this, gradient estimations are noisy but by taking samples this noise gets reduced)
- what's an activation function, draw it, and why is a ReLU better than the sigmoid
- what's the receptive field? (CNN!)
- how do I use, in my ML pipeline, the training, validation and test sets?
- why, in principle, is an autoencoder more powerful than PCA for dimensionality reduction? (non-linear vs linear)

- eigenvalue and eigenvectors, what are they, mainly what they are in our context of PCA
- ensemble learning example
- what is the cross-entropy and what is the softmax?
- draw an example of overfitting and underfitting
- what's regularization and how many approaches I have to reduce overfitting? (also ensemble learning, having more data, ...)
- what's a split function?
- VAE and GAN both use two models, use them at training time, but at test time GANs only use the generator and the VAE only uses the decoder
- draw a GAN/VAE architecture
- slack variables, geometrical interpretation (consider the margin and where they are set)