Machine Learning 2021-22
Final Exam

14 December 2021

Name: Aaron Verdaguer Gonzalez .....

Question 1: [1 point] Show that the two univariate functions e^x and $x \log x$ are convex for $x > 0$.

Hints: $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} \log x = \frac{1}{x}$, $\frac{\partial}{\partial x} x^p = p \cdot x^{p-1}$, and $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

e^x, $x > 0$

$f(x) = e^x$

$f'(x) = e^x$ Which we know is monolitically non-decreasing

$f''(x) = e^x$ Which will always be positive for $x > 0$.

x	e^x
0	1
1	2'...
2	4'...
etc.	etc...

$x \log x$, $x > 0$

$f(x) = x \log x$

$f'(x) = 1 \log x + x \cdot \frac{1}{x} = \log x + 1$

$f''(x) = \frac{1}{x}$ Which if x is never negative ($x > 0$) will always be positive

We also know that $f'(x)$ is a non decreasing function.

Question 2: [1 point] Solve the following constrained optimization problem:

$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1$$

Indicate the optimal value of x and y as well as the optimal value of the objective.

$f(x,y) = xy$

$g(x,y) = x + 2y - 1$

$\nabla f(x,y) = \lambda \nabla g(x,y)$

$2\lambda + 2\lambda = 1$

$\lambda = \frac{1}{4}$

$\frac{\partial f(x,y)}{\partial x} = y = \lambda \quad \frac{\partial g(x,y)}{\partial x} = \lambda \rightarrow y = \lambda$

$\frac{\partial f(x,y)}{\partial y} = x = \lambda \quad \frac{\partial g(x,y)}{\partial y} = \lambda 2 \rightarrow x = 2\lambda$

Optimal Values

$y = \frac{1}{4} \quad \lambda = \frac{1}{4}$

$x = \frac{1}{2} \quad \frac{1}{4} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

Question 3: [1 point] Which is the objective of k-means clustering? How does the k-means algorithm approximate a solution to this objective?

- K-means clustering is used for unsupervised learning problems. The objective is to minimize the distance between data points within a cluster. There are different types of K-means algorithms: centroid based, density based, distribution based or herarchical. Each different type tries to find possible clusters to organize the data. Centrad based recalculates centroids and clusters trying to reduce distance between data points and centroids. Other types minimize distance between cluster points differently.

Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

To use validation we have to obtain a subset, S', from our training set S, such that S' ⊂ S. Then we train an algorithm on S and evaluate it on S'. Once that is done we can repeat the process for different parameters of our model and once a model is chosen retrain the model in the whole training set. Validation loss will approximate better to true loss because the model is evaluated on unseen data which aproximates to reality.

Question 5: [1 point] Three popular models for supervised learning are decision trees, support vector machines, and neural networks. For each of these three models, list one advantage and one disadvantage.

Decision trees: they are very interpretable (help understand the problem), but very large decission trees tend to overfit.

SVM: they can create new dimensionalities and make data that was not linearly separable be linearly separable in the new space. On the other hand they are NP-hard unless we use the Kernel trick.

Neural Networks: the outcome model is very hard to interpret, but they can find solutions to very complex problems in an easier manner than we could.

Name: Aaron Verdaguer Gonzalez

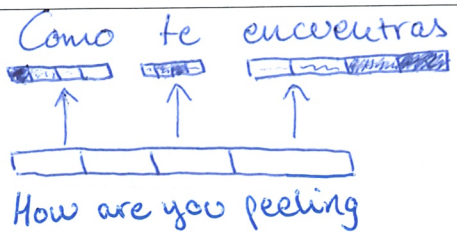
Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

The backpropagation algorithm consists on updating the weights of a neural network backpropagating the loss function into the neurons where such weights were used. In order to implement it efficiently I would suggest to use truncated backpropagation to avoid calculating your loss function on all the neurons of your neural network and all your training set. Therefore I would suggest that the most efficient way of backpropagating would be using stochastic gradient descent to avoid iterating over the whole training set and use truncated backpropagation to do it by steps.

Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

Convolutional layers apply kernels to the input data trying to focus only in important features of the input. Convolutional layers are applied in deep learning to reduce dimensionality and extract relevant information from the input or previous layer. This reduces the number of hyperparameters making the neural network more efficient. We can see each value inside a kernel as a weight and in order to apply backpropagation we have to partially derivate the loss function with the respected weight and apply what could be seen as an inverse kernel.

Question 8: [1 point] Describe qualitatively in what consists the mechanism called *attention* in deep learning and how attention is integrated in the Transformer network. You can help yourself with a diagram.



The mechanism of attention tries to highlight the most important "parts" or features of an input. In transformers we have encoder and decoder which use attention mechanisms to draw attention or to give more importance to certain parts of the input when encoding or decoding to ensure that the similarity of your query, given a value and a key focuses on the most important part of such input. Also the most common type of attention mechanism or what it is used is multi headed attention.

Name: Aaron Verdaguer Gonzalez

Question 9: [1 point] A Bayesian network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt).

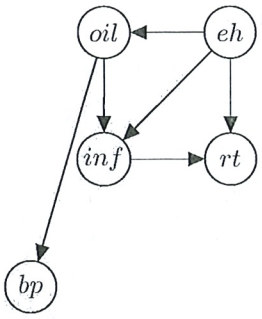


Figure 1: A simple Bayesian network

1. Write down the corresponding factorization of the joint probability distribution.
2. Use the D-Separation algorithm to determine whether the following conditional independences.
   (a) Is eh independent of bp if no evidence is provided? Why?
   (b) Is eh independent of bp if we observe that the oil is $high$? Why?
   (c) Is rt independent of bp if we observe that eh is low? Why?

1) $P(oil, eh, inf, rt, bp) = P(eh)\,P(oil|eh)\,P(inf|eh, oil)\,P(rt|eh, inf)\,P(bp|oil)$

2) a) $eh \perp\!\!\!\perp bp$: no, they are not because eh-oil-bp path is active and all paths need to be blocked. NO

b) $eh \perp\!\!\!\perp bp \mid oil$: yes because all paths are blocked due to observing oil. YES

c) $rt \perp\!\!\!\perp bp \mid eh$: no they are not because the path rt-inf-oil-bp is active. NO

Question 10: [1 point] We acquire data $\mathcal{D}$ consisting of N independent and identically distributed samples $x_i, i = 1, \ldots, N$ at different times. We assume a Gaussian generative model for $\mathcal{D}$ with constant variance σ^2 and mean determined by the output of a time-dependent model $F_\theta(t)$ with parameters θ.

1. Write down the formula for the corresponding likelihood function of θ for fixed $\mathcal{D}$.
2. Explain how Bayes theorem can be used to update the parameters θ in light of $\mathcal{D}$.

Hint: A Gaussian distributed variable y with mean μ and variance σ^2 has density $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

$P(D|\Theta) =$

2) $P(\Theta|D, \mathcal{H}) = \dfrac{P(D|\Theta, \mathcal{H})\,P(\Theta|\mathcal{H})}{P(D|\mathcal{H})}$

Basically in order to use Bayes' theorem to update the parameters Θ in light of D we need to compute the likelihood of such parameter given the data in our hypothesis space. Also we need to know how likely are parameters in our $\mathcal{H}$ which could be seen as the prior probability. And we also need to calculate how likely is data in our hypothesis space, or our evidence. Then we apply the Bayes I and we will get the posterior probability of a given parameter.

Machine Learning 2021-22
Final Exam

14 December 2021

Name: AAKASH MAROTI

Question 1: [1 point] Show that the two univariate functions e^x and $x \log x$ are convex for $x > 0$.

Hints: $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} \log x = \frac{1}{x}$, $\frac{\partial}{\partial x} x^p = p \cdot x^{p-1}$, and $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

$f(x) = e^x$; $f'(x) = e^x$

$f''(x) = e^x$; for $x > 0$; $e^x > 0$ $\therefore$ $f''(x) > 0$; e^x is convex.

$f(x) = x \log x$; $f'(x) = 1 + \log x$.

$f''(x) = \frac{1}{x}$ for $x > 0$ $f''(x) > 0$ $\therefore$ $f''(x) > 0$; $x \log x$ is conv

Question 2: [1 point] Solve the following constrained optimization problem:

$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1$$

Indicate the optimal value of x and y as well as the optimal value of the objective.

$\max_{x,y} f(x) \cdot y$ s.t. $x + 2y = 1$.

$L(f, \lambda) = xy + \lambda x + 2\lambda y - \lambda$

$\frac{dL}{dx} = y + \lambda = 0 \Rightarrow y = -\lambda$

$\frac{dL}{dy} = x + 2\lambda = 0 \Rightarrow x = -2\lambda$

putting values in constraint $x + 2y = 1 \Rightarrow -2\lambda - 2\lambda = 1 \Rightarrow \lambda = -1/4$.

$\therefore y = 1/4$ & $x = 1/2$ & $xy = 1/8$.

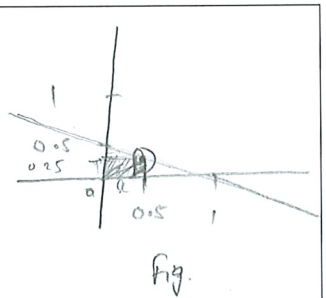
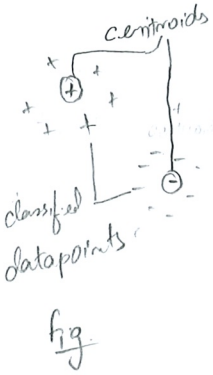


Fig.

Question 3: [1 point] Which is the objective of k-means clustering? How does the k-means algorithm approximate a solution to this objective?



fig

Objective of k means is to cluster all the given data points into n clusters. where n is a user defined number describing the no. of clusters.
The algorithm choses n arbitrary points as cluster centres (a.k.a centroids) to start with. For assignment step each ~~cluster is assigned~~ data point is assigned a cluster, whose centroid it is closest to.
For each cluster new centroid is computed by averaging all the data points in the cluster. This process of assigning clusters & calculating new centroids is repeated until convergence.
Ultimately k-means clusters the datapoints by assigning it to its nearest centroid. Kmeans calculates the euclidean distance between datapoints to get the distance.

Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

The model is trained on its ~~test~~ train set and the loss obtained is train loss. This is not an accurate estimate as it the loss is biased. A different set of data points, ones that the model wasnt trained on is used to evaluate the model's performance. This set is validation set. The model cannot overfit to the validation set. Thus by comparing model's performance on validation sets we can select better models.
the validation loss is optimistic as it influenced the model's selection. there is possiblity that validation set might be different from true data distribution. ~~Thus~~ which would result in model being selected based on its performance on data points that are not from the true distribution. Here it may perform worse on real data.

Question 5: [1 point] Three popular models for supervised learning are decision trees, support vector machines, and neural networks. For each of these three models, list one advantage and one disadvantage.

	advantage	disadvantage
SVM	Relatively simple and can run 'out of the box.'	cannot handle regression tasks. choice of kernel decides the perform, and depends on user.
Decision Tree	very simple & easily interpretable	Cannot ~~handle~~ predict real values or regression probles
Neural Networks	Universal function approximator	Require lots of data and can be stuck in local optimas.

Name: AAKASH MAROTI

Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

The algorithm ~~takes~~ propagates the gradient of loss computed at the end of a network to its nodes. It uses chain rule of derivatives to do so efficiently. where $\frac{da}{db} \times \frac{db}{dc} = \frac{da}{dc}$.

as seen in figure, we can compute gradient w.r.t. w_1 using w_2.

$$\frac{d\,loss}{dw_1} = \frac{d\,loss}{d\,w_2} \times \frac{d\,w_2}{d\,w_1}$$

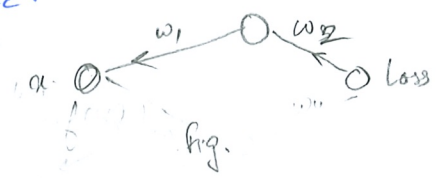


fig.

Each nodes keep record of the output & based on the signal it gets when gradients flow back from chain rule it adjusts the weights.

Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

In deep learning convolutional layers are made up of different kernels. Each kernel goes over it's input to emit a new 'image'. Several such convolution layers can help in learning different features. ~~Gradually producing high dim~~ only we now need to learn the weights of the kernel, instead of ~~all the weights~~ the weights ~~connecting all the~~ of the nodes as in a standard problem.

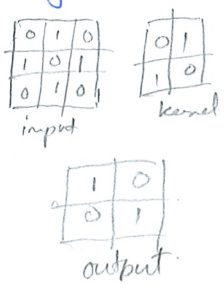


fig.

Question 8: [1 point] Describe qualitatively in what consists the mechanism called *attention* in deep learning and how attention is integrated in the Transformer network. You can help yourself with a diagram.

In deep learning attention helps a model (neural network) focus on certain parts of its inputs. It works analogous to a database query system. Where for a given query (Q) & key (K) it returns a value (V) demonstrating how much it should attend. Transformers take in entire sentences as inputs. Attention helps the network relate relevant part of the sentence with one another.

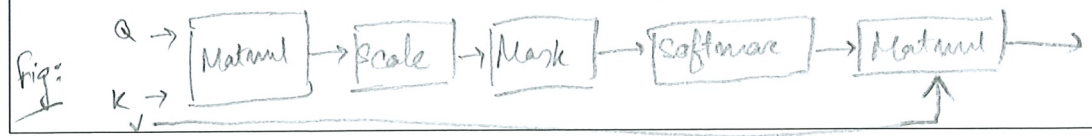


fig:

Laws have been passed making voting more difficult.

fig: Laws have been passed making voting more difficult.

Name: AAKASH MAROTI

Question 9: [1 point] A Bayesian network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt).

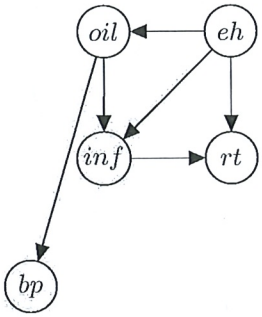


Figure 1: A simple Bayesian network

1. Write down the corresponding factorization of the joint probability distribution.
2. Use the D-Separation algorithm to determine whether the following conditional independences.
   (a) Is eh independent of bp if no evidence is provided? Why?
   (b) Is eh independent of bp if we observe that the oil is $high$? Why?
   (c) Is rt independent of bp if we observe that eh is low? Why?

2 (a) No, ~~EH is indep of BP~~ eh depends on bp because no evidence of oil is provided
(b) Yes, eh is indepent of bp because intermediate node (i.e. oil)'s evidence is provided.
(c) Yes, given no evidence on oil eh is dependent on bp, thus influencing rt.
1. It can be factored into $P(eh)\ P(oil|eh)\ P(bp|oil)\ P(rt|eh,inf)$.

Question 10: [1 point] We acquire data $\mathcal{D}$ consisting of N independent and identically distributed samples $x_i, i = 1, \ldots, N$ at different times. We assume a Gaussian generative model for $\mathcal{D}$ with constant variance σ^2 and mean determined by the output of a time-dependent model $F_\theta(t)$ with parameters θ.

1. Write down the formula for the corresponding likelihood function of θ for fixed $\mathcal{D}$.
2. Explain how Bayes theorem can be used to update the parameters θ in light of $\mathcal{D}$.

Hint: A Gaussian distributed variable y with mean μ and variance σ^2 has density $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

$P(\theta|D,H) = \frac{P(D|\theta,H)\,P(\theta|H)}{P(D|H)}$ $\qquad P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$

1. $P(\theta, D, F) = \frac{P(D|\theta,F)\,P(\theta|F)}{P(D|F)}$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Marian Iglesias Blanco

Question 1: [1 point] Show that the two univariate functions e^x and $x \log x$ are convex for $x > 0$.

Hints: $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} \log x = \frac{1}{x}$, $\frac{\partial}{\partial x} x^p = p \cdot x^{p-1}$, and $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

For an univariate function, if $f''(x) > 0$ the function will be convex. Derivate to prove it:

$f(x) = e^x$

$f'(x) = e^x$

$f''(x) = e^x$

↳ for $x > 0$ it will hold $f''(x) > 0$, hence it is convex.

$f(x) = x \log x$

$f'(x) = \log x + x \cdot \frac{1}{x}$

$f''(x) = \frac{1}{x}$

↳ for $x > 0$ it will hold $f''(x) > 0$, hence it is a convex function.

Question 2: [1 point] Solve the following constrained optimization problem:

$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1$$

Indicate the optimal value of x and y as well as the optimal value of the objective.

Since the constraint is an equality, we use the Lagrangian conditions to maximize xy and holding the constraint.

$\mathcal{L}(x,y,\lambda) = xy + \lambda(1 - x - 2y)$

Lagrangian conditions → 1) $\frac{\partial \mathcal{L}(x,y,\lambda)}{\partial x} = 0$ 2) $\frac{\partial \mathcal{L}(x,y,\lambda)}{\partial y} = 0$

↳ Partial derivative set to 0.

→ Procedure in the additional sheet!

Found:

$$\begin{cases} x = 1/2 \\ y = 1/4 \end{cases}$$

(*) A) Find x_i which minimize distance to actual centroids μ_k and assign the closest

B) Find centroid which minimize distance with the datapoints of the belonging cluster for each k cluster.

Question 3: [1 point] Which is the objective of k-means clustering? How does the k-means algorithm approximate a solution to this objective?

(infer a pattern without any label)

In k-means we want to summarize original data with a single categorical value, so grup data in k clusters. k-means approximate the solution in the following way:

1. Initialize with random k-samples as centroids
2. Iterate until convergence.
   - (*) A) - Assign each datapoint to the closest centroid
   - B) - Recompute centroids based on new cluster.

→ It is a centroid-based clustering that will find k clusters without any target provided.

Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

Validation consist on training the model on a dataset S, and once trained, used the model to predict outcomes of an unseen dataset V. However, with this we compare models using save S for (as V for testing) training, and hence maybe one model could just perform better because of this, for that reason it is still optimistic, and it would be better to use cross-validation, so you get an average performance changing validation and training set, obtaining a better approximation of true loss.

Since S is sampled from D, and you train your model assuming same distribution.

Question 5: [1 point] Three popular models for supervised learning are decision trees, support vector machines, and neural networks. For each of these three models, list one advantage and one disadvantage.

Decision trees
- Easy to interpret
- Greedy (≠ tree depending on S)

SVM
- Not easy to interpret the obtained results
- Besides minimizing classification loss, maximizes the geometrical margin between classes (more robust)

NN
- Lot of engineering (hyperparameter selection)
- End-to-end model

Name: .....

Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

Backpropagation consists on:

1) Computing the gradient of $L_S(w)$ with respect to w

2) Update of weights based on gradient descent (or stochastic) method.

to implement it efficiently, vectorization can be done for acyclic networks, meaning an update weight matrix ΔW_{ij} can be obtained and applied allowing parallel update being efficiently implemented. * this matrix depends on: the learning rate, the error $(y_k - x_k)$ and the gradient to each element of the matrix

Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

Convolutional layer is composed by filters, since CNN are based on convolution of these filters with input. The convolutional layer is defined by a number of filters with $n \times m$ size, with the possibility to specify padding and stride of how the filter is implemented.

The NN will learn the weights defining those filters, so weights are shared for different inputs (since it is convolved). This makes that now backpropagation is also a filter, and convolution is done to obtain weights update.

Question 8: [1 point] Describe qualitatively in what consists the mechanism called *attention* in deep learning and how attention is integrated in the Transformer network. You can help yourself with a diagram.

Attention in DL is modelled as the retrieval of v_i for a query q from a key k_i in a database. So, similarity between q and k_i is measured, and from this and attention coefficient is computed, and a weighted combination for v_i is returned.

transformer network does not use recurrency because attention mechanism is implemented, multi-head attention is done, so in the first layer we have pairwise embeddings. in the second one pairs of pairwise embeddings, and so on.

Name: .....

Question 9: [1 point] A Bayesian network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt).

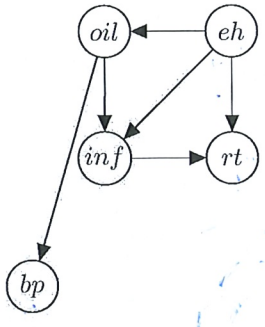


Figure 1: A simple Bayesian network

1. Write down the corresponding factorization of the joint probability distribution.
2. Use the D-Separation algorithm to determine whether the following conditional independences.
   (a) Is eh independent of bp if no evidence is provided? Why?
   (b) Is eh independent of bp if we observe that the oil is $high$? Why?
   (c) Is rt independent of bp if we observe that eh is low? Why?

(2) a) No, they will be dependent since oil node is blocking info flow. It is a head-to-tail node.

b) Yes, now they are independent because oil is provided, so eh can be explained by oil and bp too.

c) No, because although we observe eh, oil node into is not being observed, so info flow is being blocked and they are dependent.

Question 10: [1 point] We acquire data $\mathcal{D}$ consisting of N independent and identically distributed samples $x_i, i = 1, \ldots, N$ at different times. We assume a Gaussian generative model for $\mathcal{D}$ with constant variance σ^2 and mean determined by the output of a time-dependent model $F_\theta(t)$ with parameters θ.

1. Write down the formula for the corresponding likelihood function of θ for fixed $\mathcal{D}$.
2. Explain how Bayes theorem can be used to update the parameters θ in light of $\mathcal{D}$.

Hint: A Gaussian distributed variable y with mean μ and variance σ^2 has density $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

Procedure in a sheet apart.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Marian Iglesias

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Question 2

Lagrangian conditions → set Lagrangian to 0

$\max_{x,y} xy \rightarrow \frac{\partial \mathcal{L}}{\partial x} = 0$ and $\frac{\partial \mathcal{L}}{\partial y} = 0$

$\nabla_{\lambda} g(\lambda) = 0$.

$\mathcal{L}(x,y,\lambda) = xy + \lambda(1 - x - 2y)$

$\frac{\partial \mathcal{L}}{\partial x} = y - \lambda$ $\quad$ $\frac{\partial \mathcal{L}}{\partial y} = x - 2\lambda$

$\boxed{\lambda = y^*}$ $\quad$ $\boxed{2\lambda = x^*}$

$g(\lambda) = 2\lambda^2 + \lambda(1 - 2\lambda - 2\lambda) = 2\lambda^2 + \lambda(1 - 4\lambda)$

$\boxed{g(\lambda) = 2\lambda^2 + \lambda - 4\lambda^2 = \lambda - 2\lambda^2}$

$\frac{\nabla g(\lambda)}{\partial \lambda} = 1 - 4\lambda \rightarrow \boxed{\lambda = \frac{1}{4}}$

$\hookrightarrow \boxed{x = 2 \cdot \frac{1}{4} \text{ and } y = \frac{1}{4}}$

$x = 1/2$

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Marian Iglesias

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Question 10

Each datapoint comes from a gaussian distribution

$$p(x_i|\phi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{where } \mu = f_\phi(t)$$

Since data is i.id. for each t:

$$p(D|\phi) = \prod_{i=1}^{N} p(x_i|\phi)$$

$$p(D|\phi) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$p(D|\phi) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

apply log to get rid of $\prod$

$$\log(p(D|\phi)) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N + \sum_{i=1}^{N} -\frac{1}{2}\left(\frac{x - f_\phi(t)}{\sigma}\right)^2$$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Eduard Alcobé Garcia

Question 1: [1 point] Show that the two univariate functions e^x and $x \log x$ are convex for $x > 0$.
Hints: $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} \log x = \frac{1}{x}$, $\frac{\partial}{\partial x} x^p = p \cdot x^{p-1}$, and $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

f is convex → f' is non decreasing → f'' is positive

For e^x: $f(x) = e^x$

$f'(x) = e^x$

$f''(x) = e^x$ For $x > 0$ this is positive so e^x is a convex function

For $x \log(x)$ $f(x) = x \log x$

$f'(x) = \frac{\partial f(x)}{\partial x} = \log x + x \frac{1}{x} = \log x + 1$

$f''(x) = \frac{\partial f'(x)}{\partial x} = \frac{1}{x}$ which is positive for $x > 0$ so this is a convex function

Question 2: [1 point] Solve the following constrained optimization problem:

$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1$$

Indicate the optimal value of x and y as well as the optimal value of the objective.

We construct the lagrangian

$L(x, y, \lambda) = xy + \lambda(1 - x - 2y)$

We compute the gradient and equalize to 0

$\nabla_{x,y} L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} y - \lambda \\ x - 2\lambda \end{pmatrix}$ now $\nabla_{x,y} L = 0 \rightarrow$ $y - \lambda = 0 \rightarrow y = \lambda$, $x - 2\lambda = 0 \rightarrow x = 2\lambda$

So, the dual $g(\lambda)$ when ~~substituting~~ replacing x and y in terms of λ is

$g(\lambda) = 2\lambda^2 + \lambda(1 - 2\lambda - 2\lambda) = 2\lambda^2 + \lambda - 4\lambda^2 = -2\lambda^2 + \lambda$

We compute the gradient of the dual to find the best parameter λ

$\nabla_\lambda g(\lambda) = -4\lambda + 1$ $\nabla_\lambda g(\lambda) = 0 \rightarrow -4\lambda + 1 = 0 \rightarrow \lambda^* = \frac{1}{4}$

So as $x = 2\lambda = 2\frac{1}{4} = \frac{1}{2}$ and $y = \lambda = \frac{1}{4}$

$x = \frac{1}{2}$ $y = \frac{1}{4}$ $\lambda^* = \frac{1}{4}$

Question 3: [1 point] Which is the objective of k-means clustering? How does the k-means algorithm approximate a solution to this objective?

The objective of k-means clustering is to find K clusters, so a group of data according to structure in data as it's an unsupervised learning algorithm. It is centroid-based, meaning that the algorithm minimize the distance between the centroid of a cluster and data points which form the cluster. K-means algorithm receives as inputs the training set $S=(x_1,\dots x_m)$ and an integer K. Then randomly K centroids are selected. In order to approximate a solution to the objective, clusters are recompute given the initial centroids, then centroids are recomputed given those clusters and these operations are repeated until convergence.

Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

Validation is a technique to overcome overfitting. Basically the data set is splitted in training set S and validation set V. For example let's say 80% S and 20% V. Then the model is trained using S but the error is computed using V so we get a validation loss which will be more likely a good estimation for the true loss than the training loss. ~~However~~ ~~This~~ Validation is usually used to perform model selection and compute the error between these models.
However the validation loss still optimistic because fewer data was used for training so less overfitting is likely to occur. Furthermore the size of the validation set may not be representative.

Question 5: [1 point] Three popular models for supervised learning are decision trees, support vector machines, and neural networks. For each of these three models, list one advantage and one disadvantage.

Decision trees are very easy to understand, and to interpret, and to express knowledge of the problem ~~to the~~ the user try to solve. However, as disadvantage, it is greedy and tends to ~~ove~~ suffer overfitting.

SVM is more complex as a transformation from the initial space to another to linearly separate data using inner product computed with Kernels. Also, it allows mislabelled examples as it has soft margin. Nevertheless, it has disadvantages as it is very dependent to kernel parameters and soft margin constant

Neural networks are great to solve non linear function problems because NN are based on activation function which give flexibility. Nevertheless is difficult to understand what is happening so the result model is hard to interpret

Name: Eduard Alcalá García

Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

The backpropagation algorithm consist in computing the gradient with respect to the weights and then use gradient descend to update the weights of a given problem. As the name indicates it goes backwards (from the output to the input) As it can be difficult to use the gradient due to memory issues, a partial gradient applied to a subset may be a good option. Also truncated backpropagation in the case of RNN can be considered to avoid memory problems. Another important aspect is to use matrix operations to avoid computational blowup. Depending on the neural network (RNN) gradient clipping or architectural change to avoid gradient vanishing or explosion are required. Finally, it is important to highlight that the convergence of this algorithm is slow and that a lot of data is required and may suffer overfitting. So in order to decide which is the best strategy (partial gradient, gradient clipping...) the user may take these aspects into account as well as the neural network type of course.

Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

Convolutional layers are basically filters which are applied to the whole image, so the pattern is repeated. Convolutional layers are not fully connected and share ~~dot~~ weights. When applying a convolutional layer or a filter the image shrinks, however the space features are maintained. They are very helpful to reduce the number of parameters to consider. In practice convolutional layer are small matrix (filters) that are combined with subsampling layers which reduces the number of nodes using max pooling: the higher input is kept $0 = max_i$. So CNN is formed by convolutional layers and subsampling layers and usually a full connected layer at the end.

To apply backpropagation we have to be concious that it's also ~~for~~ a convolutional problem. In order to discover the best filters to apply, the backpropagation is used, but we have to taken into consideration that all is based on matrix and partial gradients across features.

Question 8: [1 point] Describe qualitatively in what consists the mechanism called *attention* in deep learning and how attention is integrated in the Transformer network. You can help yourself with a diagram.

The attention mechanism consist in having memory of relationships between elements. Without using recurrency in the case of Transformers and with recurrency in RNN. So basically attention is the ability of an algorithm to keep information about previous iterations/states... so it has memory.
In deep learning this is achieved using RNNs and gates such as LSTM. However in Transformers it is based on encoder-decoder approach. Both parts encoder and decoder has attention by itselfs and also there is attention between encoder and decoder blocks. So the information flows and in fact the output of the decoder is then used as additional input in the decorder in order to record the events.
The attention mechanism is essential for language model applications.

Name: Edward Alcahé Garcia

Question 9: [1 point] A Bayesian network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt).

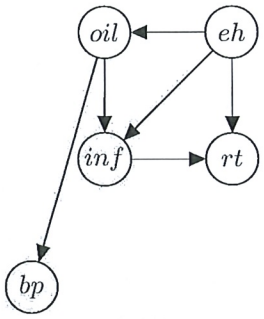


Figure 1: A simple Bayesian network

1. Write down the corresponding factorization of the joint probability distribution.
2. Use the D-Separation algorithm to determine whether the following conditional independences.
   (a) Is eh independent of bp if no evidence is provided? Why? No
   (b) Is eh independent of bp if we observe that the oil is $high$? Why? Yes
   (c) Is rt independent of bp if we observe that eh is low? Why? No

1. $P(oil, inf, eh, bp, rt) = P(eh)\, P(oil|eh)\, P(bp|oil)\, P(inf|eh,oil)\, P(rt|eh,inf)$

2. a) $eh \perp\!\!\!\perp bp$? The path $eh \rightarrow oil \rightarrow bp$ is not blocked as it is a head to tail without any knowledge so eh and bp are not independent $eh \not\perp\!\!\!\perp bp$

   b) $eh \perp\!\!\!\perp bp \mid oil$? All paths are blocked in oil as a central node because going directly from $eh \rightarrow oil \rightarrow bp$ is a head to tail knowing the central node and going $eh \rightarrow inf \leftarrow oil \rightarrow bp$ there is a tail to tail in oil knowing the central node. So yes $eh \perp\!\!\!\perp bp \mid oil$

   c) $rt \perp\!\!\!\perp bp | eh$? Path $rt \leftarrow inf \leftarrow oil \rightarrow bp$ is openned as there is a head to tail with inf as central node and a tail to tail with oil as central node without any knowledge. So, NO $rt \not\perp\!\!\!\perp bp | eh$

Question 10: [1 point] We acquire data $\mathcal{D}$ consisting of N independent and identically distributed samples $x_i, i = 1, \ldots, N$ at different times. We assume a Gaussian generative model for $\mathcal{D}$ with constant variance σ^2 and mean determined by the output of a time-dependent model $F_\theta(t)$ with parameters θ.

1. Write down the formula for the corresponding likelihood function of θ for fixed $\mathcal{D}$.
2. Explain how Bayes theorem can be used to update the parameters θ in light of $\mathcal{D}$.

Hint: A Gaussian distributed variable y with mean μ and variance σ^2 has density $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

1. $P(D|\theta, H) = \theta \cdot \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - f_\theta(t)}{\sigma}\right)^2\right) dx$

2. $P(\theta|D,H) = \frac{P(D|\theta,H)\, P(\theta|H)}{P(D|H)}$ So Bayes theorem can be used to know the best θ calculating the posterior of θ given D knowing the likelihood of θ, the prior of θ and the evidence of D. Comparing the posterior probability of each θ we ~~may be at~~ can update the parameter θ.

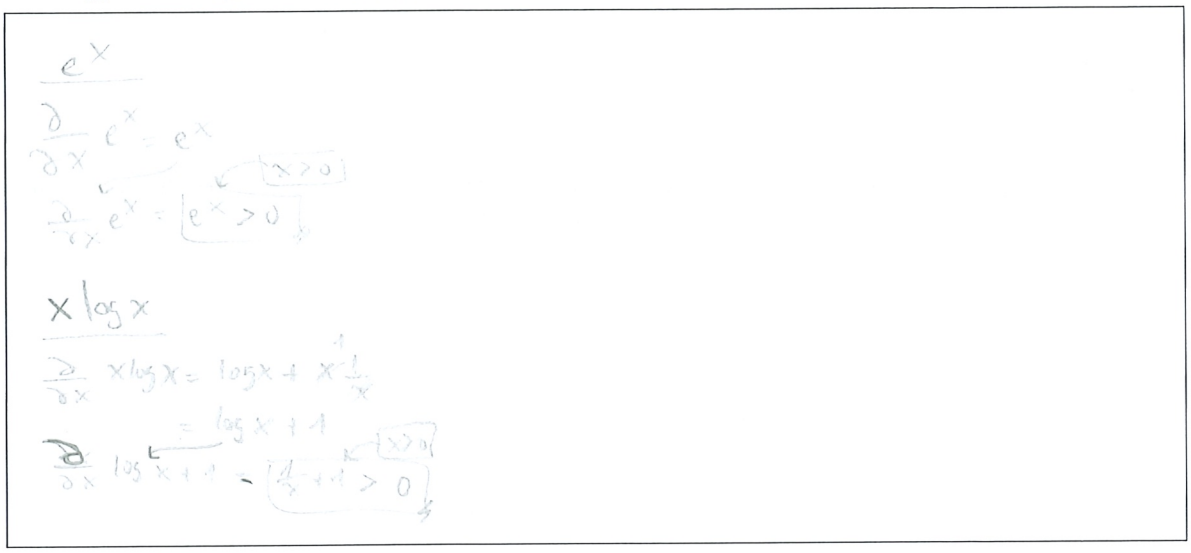
Machine Learning 2021-22
Final Exam

14 December 2021

Name: BRUNO FERNANDO SICUA PLATA

Question 1: [1 point] Show that the two univariate functions e^x and $x \log x$ are convex for $x > 0$.

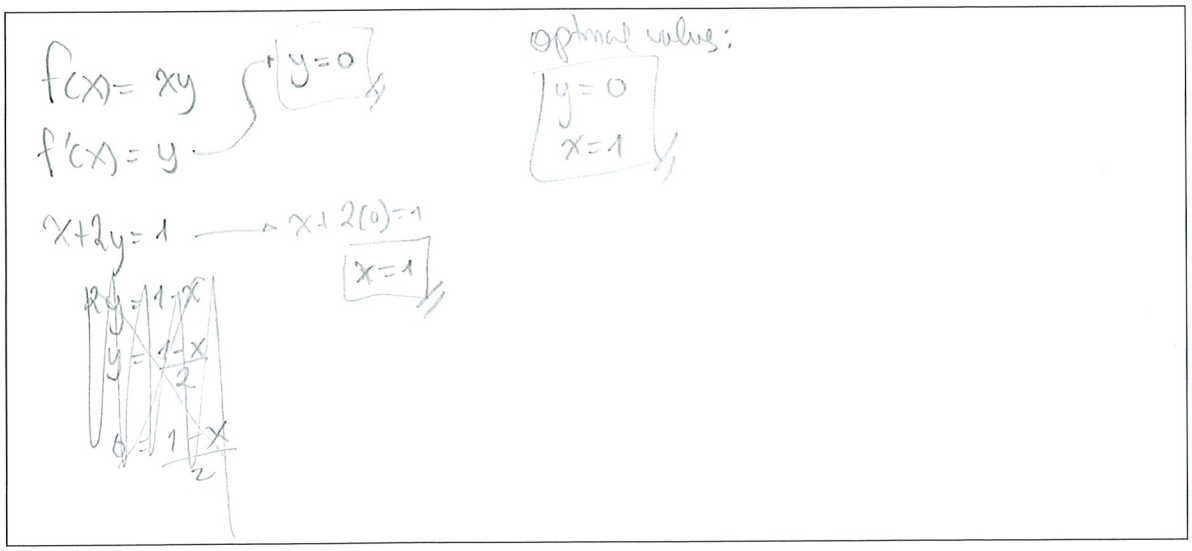
Hints: $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} \log x = \frac{1}{x}$, $\frac{\partial}{\partial x} x^p = p \cdot x^{p-1}$, and $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.



Question 2: [1 point] Solve the following constrained optimization problem:

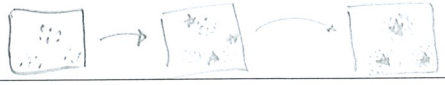
$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1$$

Indicate the optimal value of x and y as well as the optimal value of the objective.



Question 3: [1 point] Which is the objective of k-means clustering? How does the k-means algorithm approximate a solution to this objective?

K-means is an Unsupervised learning algorithm, this algorithm's goal is clustering.
The way how it works is: First we give a value for k (number of clusters that we want to find in our data) then k means will initialize k centroids, this centroids will move towards data based on the distance. This process will happen until each k centroid will not change, so they are finally centroids of k clusters.



Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

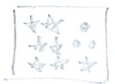
Validation can be used in the following way: we can get our data and separate it in one set of traing, other for testing and other for validation so we will have:

At the beginning we will train the model with the training set, then we will verify the performance with data that the model "have never saw" so we will use testing set...

Continue in extra paper

Question 5: [1 point] Three popular models for supervised learning are decision trees, support vector machines, and neural networks. For each of these three models, list one advantage and one disadvantage.

Decision Trees

This model is very useful and easy in certain escenarios, If we have some data like  we can find with Decision Trees the boundaries that can do a good classification. But this is not always what happens in some occasions Decision Tree can fail in order to do a good classification, so it will fail in overfitting, this could happen if the Decision Tree have a lot of levels, so it will overfit.

Continue in paper

Name: BRUNO FERNANDO SILVA PLATA

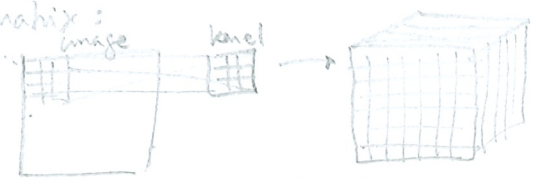
Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

When we use Feedforward Neural Network we train the network going to the right:

Once we obtain our prediction we check how similar it is with the real target so we will see (if there are) an error this error will help us to improve our weights. But if we want to improve the weights we need to go backwards, so... [continue paper]

Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

The idea of convolutional layer is to apply a kernel in a matrix in order to capture relevant features. This can be implemented in the following way: we implement a layer (a conv. layer) that define a kernel size, and the output channel, this will go through the matrix (image representation) and it will give us a new matrix:



This process also needs a stride and a padding to be defined. Sometimes it is followed by a pooling. [Continue in other paper]

Question 8: [1 point] Describe qualitatively in what consists the mechanism called *attention* in deep learning and how attention is integrated in the Transformer network. You can help yourself with a diagram.

Attention, as the name says, give some attention to a certain value. The idea is like softmax in NN where at the end, the prediction is based on the implementation of softmax in the final vector this will give us the highest score represented by a high probability, because softmax will make the value very high because it is the value that we should focus. This idea is like attention, the "area" where we can find attention is with sequential data... [continue paper]

Name: BRUNO FERNANDO SILVA PLATA

Question 9: [1 point] A Bayesian network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt).

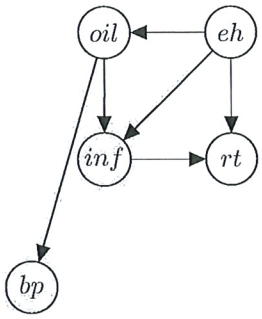


Figure 1: A simple Bayesian network

1. Write down the corresponding factorization of the joint probability distribution.
2. Use the D-Separation algorithm to determine whether the following conditional independences.
   (a) Is eh independent of bp if no evidence is provided? Why?
   (b) Is eh independent of bp if we observe that the oil is $high$? Why?
   (c) Is rt independent of bp if we observe that eh is low? Why?

① $= p(bp|oil)\,p(oil|eh)\,p(eh)\,p(inf|oil,eh)\,p(rt|inf,eh)$

② (a) No, because there is no block in the path {eh, oil, bp}

(b) Yes, because there is a "block" in the path, because of oil

(c) No, because there is no "block" in {bp, oil, inf, rt}

Question 10: [1 point] We acquire data $\mathcal{D}$ consisting of N independent and identically distributed samples $x_i, i = 1, \ldots, N$ at different times. We assume a Gaussian generative model for $\mathcal{D}$ with constant variance σ^2 and mean determined by the output of a time-dependent model $F_\theta(t)$ with parameters $\boldsymbol{\theta}$.

1. Write down the formula for the corresponding likelihood function of $\boldsymbol{\theta}$ for fixed $\mathcal{D}$.
2. Explain how Bayes theorem can be used to update the parameters $\boldsymbol{\theta}$ in light of $\mathcal{D}$.

Hint: A Gaussian distributed variable y with mean μ and variance σ^2 has density $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

① $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$ ← likelihood, ← prior, ← data

So: the likelihood function of θ for fixed D is $P(D|\theta)$

② Bayes theorem can be used to update the parameters θ using the likelihood $P(D|\theta)$ the prior that in this case is $P(\theta)$ and the Data that is determined by the gaussian generative model.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: BRUNO FERNANDO SILVA PLATA

Assignatura: Grup: Curs:

Centre/Estudi:

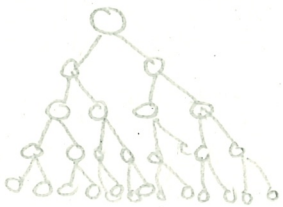
Professor/a: ANDERS JONSSON & VINÇEC GOMEZ

Data:

(4) ... So then we will have a more realistic peformance, but we can continue doing one last check, we can use validation set. This procedure will give us an optimistic estimate of the true loss because we are using data that the model have never saw, but also we are checking the model performance with 2 stages.

Validation can also be used when we want to do CROSS-VALIDATION in this technique (a very recommended one) we will split the data in K pieces, then we will run the model using some pieces as training and the other for testing, this with all the experiments, then we will average the performance in order to see the model performance this is a good and optimistic estimate of the true loss

(5) Decision Tree with many levels, will fail because of overfitting



Support Vector Machine

This model is very useful because can help to classify between classes with a ~~trick~~ trick: change dimension. This trick is kernel trick, where we can find an hyperplane:

However, in some cases SVM will fail, for example when we have data like:

in this case SVM will not classify as well as the Decision tree for example. Because of the hyperplane, and it will be costly in comparision with D.T.

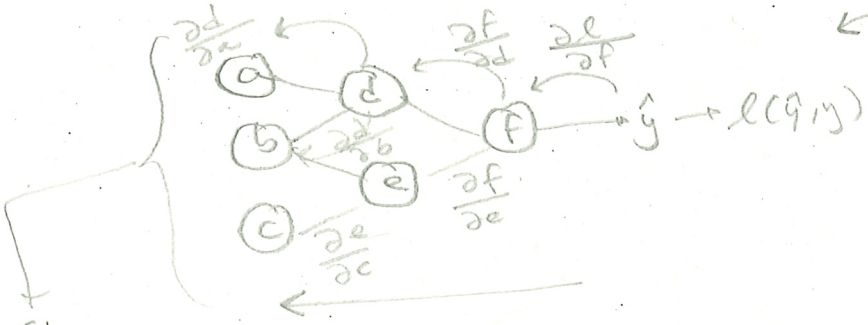
Neural Networks

NN are very useful because help us to find features in our data this because of the hidden layers that use, so in tasks with unstructured data will be very helpful. But in other tasks will not be the best option because of it complexity. For example with a structured data maybe will do a good job finding the best decision boundary but it will costly in comparision with other models. Also NN can have some "intern" problems like vanishing gradient, between the activation in the weights and problems related to the overfitting that can happen because of the size of the architecture.

---

(6) ... we will use backpropagation in order to modify the weights based on the performance that the NN had before. This will be done using chain rule:

EXAMPLE



Changes

in a: $\frac{\partial f}{\partial d}\frac{\partial d}{\partial a}$

in b: $\frac{\partial f}{\partial d}\frac{\partial d}{\partial b} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial b}$

in c: $\frac{\partial f}{\partial e}\frac{\partial e}{\partial c}$

(8) ... In sequential data we have cases where if we use RNN, for example, this will forget valuable information because of the lacke of "attention" to the important data. But in more advanced arquitectures we can use attention in order to do even more diffrent tasks. In Transformers for example we can use attention in order to perform tasks that involve encoders and decoders, this is very helpful in tasks like language translation. The transformers orchitecture is like the following diagramm:

Universitat Pompeu Fabra Barcelona

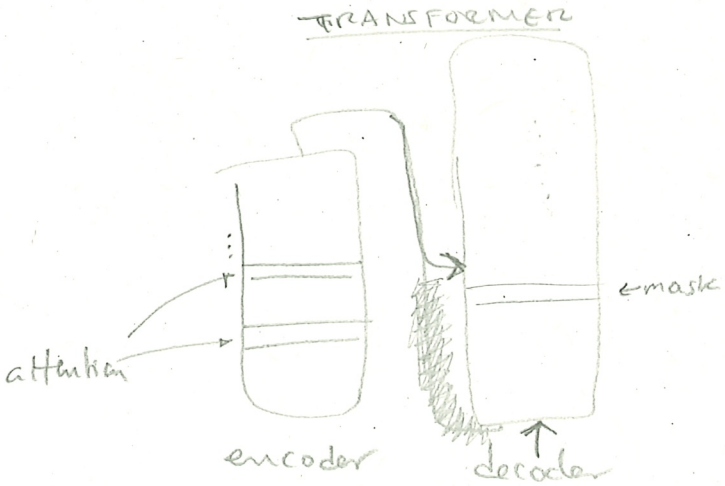
Nom i cognoms: BRUNO FERNANDO SILVA PLATA

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a: ANDERS JONSSON & VINÇEC GOMEZ

Data:



The architecture is more robust, and use in many steps the attention layer. At the moment when data is in the encoder it use attention because it will help to give encoder the power of realize some independant tasks like prediction of the next word. Independant in the way that it will not use the decoder. An example can be Bert. The decoder can also be considered in an individual way.

(7) In convolutional layers we have our weights in the kernel matrix that is doing calculations through the matrix. Because it contains the weights, the backpropagation will update these values.

$$P(D|\theta) = \frac{P(\theta|D)\,P(D)}{P(\theta)}$$

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$