Machine Learning 2021-22
Final Exam

14 December 2021

Name: ..David...Cáceres...Cruz.....

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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

1) $f(x) = e^x$, $f'(x) = e^x$, $f''(x) = e^x$ → is positive for all x

This means it is convex.

$f'(x)$ → non-decreasing, $f''(x)$ → positive

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

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

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

we can also see here that it is convex.

For $x > 0$ $f'(x)$ is non-decreasing while $f''(x)$ is positive.

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.

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

$\nabla L(x,y,\lambda) = \begin{pmatrix} \nabla_x L(x,y,\lambda) \\ \nabla_y L(x,y,\lambda) \\ \nabla_\lambda L(x,y,\lambda) \end{pmatrix} = \begin{pmatrix} y + \lambda \\ x + 2\lambda \\ x + 2y - 1 \end{pmatrix} = 0$

$\begin{cases} y = -\lambda, \\ x = -2\lambda \end{cases} \rightarrow -2\lambda + 2(-\lambda) - 1 = 0$

$-2\lambda - 2\lambda - 1 = 0$

$\lambda^* = -1/4$

$y^* = 1/4$ $x^* = 2/4$

~~$y^* = 1/5$~~ ~~$x^* = 2/5$~~

If we substitute $L_o = \frac{1}{4} \cdot \frac{2}{4} + \underbrace{\left(-\frac{1}{4}\right)\left(\frac{2}{4} + \frac{2}{4} - 1\right)}_{0}$

$L_o = \frac{1}{2}$

$\max xy \rightarrow \frac{1}{4} \cdot \frac{2}{4} = \frac{2}{4} = \frac{1}{2}$

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 is to group N data points into K clusters according to the distance to K centroids. To do this, the algorithm randomly initializes K clusters at random coordinates (K is chosen by the learner). Each data point in the space is assigned by proximity to one centroid. Once we have the clusters, we recompute the centroids as the mean of all the point coordinates in each cluster. We reassign each data point to the closest centroid calculating the distances. This process is repeated iteratively until no more changes are achieved. The shape of the clusters correspond to Voronoi cells. They cannot find more complicated cluster shapes.

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 perform validation, we separate a subset of the training set (usually 10 - 20%). The model, during the training process will be evaluated in this new set every certain number of epochs. This way we can know if the model is able to generalize to unseen samples and track the progress of training. It is useful to select the optimal parameters which will allow the model to generalize better. However, it is still optimistic, as it is a small subset of the training set. Depending on how was this data collected, the real distribution can be very different. So it is subject to the quality of the dataset used.

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

- Advantage: they are easy to interpret
- Disadvantage: they tend to overfit to the data

SVM

- Advantage: They generalize the perceptron (take into account non-linearity)
- Disadvantage: They are highly dependent on the parameters (e.g. C, kernel constants, etc.)

Neural networks

- Advantage: they can represent any hypothesis function $h(x)$
- Disadvantage: they are more computationally expensive. They can also converge to local minimum.

Name: Daniel Encades Gault

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

The backpropagation algorithm is the algorithm used to update the weights in neural networks. Once the input has been processed through all the layers, the loss function is calculated. Backpropagation updates all the weights in the backward direction in a similar way to gradient descent. Every weight is updated according to the gradient of the loss function with respect to that weight. This gradient is multiplied by an hyperparameter (η → learning rate), which determines how fast we descend the gradient. We must select a high η not causing divergence and lower the value when reaching the plateau.
The principal disadvantages are that it can converge to local minimum, it is computationally expensive, among others.

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?

A convolutional layer is composed of several filters which are convolved by the entire image, extracting features. Each of these filters is composed of weights, and are learned in a similar way to the weights in FFN. In contrast with these last, CNNs are not fully connected and share the weights. Several convolutional layers with subsampling layers and activations are concatenated to extract low level features (lines, etc.) to high level features (e.g. faces, buildings, etc.)
To apply backpropagation, we need to do convolutions (similar to the ones done in forward direction) This convolutions are done by filters containing this time the gradients with respect to the weights.

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 refers to a mechanism to evaluate which parts of the input are important to predict the output. This attention is integrated in the Transformer network by calculating similarities between different parts of the input/output. For example, if we want to translate an English sentence to Spanish, we would introduce the English sentence to the encoder of the Transformer network. The multi-headed attention block will generate a vector of similarities of each word of the sentence with each other. Similarities are computed → $S(Q,K) = QK^T$ and normalized through a softmax function

Example. The exam is difficult

- 0.1 → similarity with 'The'
- 0.8 → 'exam'
- 0.3 → 'is'
- 1 → 'difficult'

→ when taking "difficult" as input we will have the context of "exam".

Name: Daniel Cardona Géret

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).

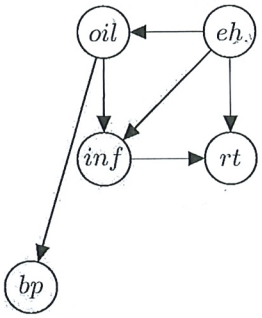


Figure 1: A simple Bayesian network

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?

(a) No, as they form a head-to-tail network. When there is no information of oil, they are dependent. (eh) → (oil) → (bp)

(b) In this case, in contrast with a, we observe oil, so bp can be fully explained by oil. They are independent.

(c) The evidence of "eh" blocks one of the paths from "rt" to "bp". However, there is another path (bp → oil → inf → rt) which is not blocked. They are still dependent.

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

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

$p(\theta^*|D) = \underset{\theta}{\operatorname{argmax}} \prod_i^N p(x_i|\theta)$ → likelihood (max.)

2. The optimal parameters are found with the posterior of θ, $p(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$ (Bayes I)
Multiplying the likelihood $P(D|\theta)$ (probability of having generated D having θ) by the prior of the parameters. It is then divided by the probability of the data. In this case, the probability of the data is modeled with the density function that generates the points. This give us the optimal parameters given the data D.

Machine Learning 2021-22
Final Exam

14 December 2021

Name: MIRIAM CARVAJAL RODRÍGUEZ

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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

To know if a function $f(x)$ is convex we can know it by the relation of $f(x)$ convex $\equiv$ $f''(x)$ is non-negative
we are going to calculate $f''(x)$ for each one (for $x>0$). For $f(x)=e^x$:

1st derivative $\frac{\partial e^x}{\partial x} = e^x$ → 2nd $\frac{\partial (e^x)'}{\partial x} = (e^x)' = e^x$ → for $x>0$ → $f''(x) = e^x$ → $f''(x) > 0$
so $f(x)=e^x$ is convex for $x>0$

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

1st derivative $\frac{\partial}{\partial x}(x \log x) = \frac{\partial x}{\partial x} \log x + x \frac{\partial (\log x)}{\partial x} = \log x + x \cdot \frac{1}{x} = \log x + 1$

2nd derivative $\frac{\partial}{\partial x}(\log x + 1) = \frac{1}{x} + x$ for $x>0$ we have that the second derivative always $f''(x>0) > 0$ → $f(x) = x \log x$ is convex for $x>0$

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.

Function we want to optimize → $f(x,y) = \max_{x,y} xy$
but we must consider the range between $x + 2y = 1$ → constrained.
Our function is then ↳ $x = 1 - 2y$

$$f(x,y) = \max_{x,y} xy = \max_{x,y} (1-2y)y = \max_{y} y - 2y^2$$

EXERCISE DONE IN THE ~~PAGE~~ SHEET

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

OBJECTIVE → is to agroup the data points in a data set, by their similarity. This similarity is given by the spatial distribution of the points, keeping in the same cluster points that are close between them. This works for an initial specified number k of clusters. The algorithm does the following: → by its distance to centroids (of each cluster)

1.- User specifies the desired number k of clusters
2.- Randomly initializates the centroids of the clusters
3.- Asign each data point to a cluster (to the more close centroid)
4.- Re-compute the centroids with the new cluster (mean point between all in the cluster)
5.- Iterate this actualization of the clusters for T iterations from the point 3. (o stop before if converges)

Take into consideration that this clustering is used for unsupervised learning problems, since data is not labeled. So we will need help from an expert to see the meaning of agrupations. Also the random initialization can give us a different clustering if we repeat the experiment.

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 used for supervised learning problems in order to test the model in front of unseen data for the training, In practice what we do is to split the given data into a training set (S) and a validation set (V).

Validation loss is more similar to the true loss, because it describes how the model behaves in front of "unseen" points (same in the case of true loss) and have an idea about if the model is overfitted or not. → Because the model is trained to reduce the training loss,

To perform a model selection we can use validation:
1.- select the num to test.
2.- Train each model with the training test
3.- Perform the validation: compute validation loss (test the validation set)
4: Compare the validation loss among models and select the model with smaller loss

I CONTINUE IT IN THE SHEET

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 TREE
- Adv. → good interpretability of the model, it is easy to understand and explain.
- DIS. → Data dependent model, the model can be overfitted.

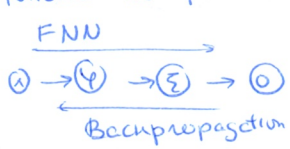
SVM
- Adv. → allows us to do non-linear data separation (by increasing dimensionality and convex function, would reach the minimum. hyperplane separation).
- DIS. → This dimensionality increasement causes an increasement in the computational cost.

NN
- Adv → can be used for feature extraction, so it does not required specified features and it is not dependent to them.
- DIS → requires a lot of data for its training.

Name: Miriam Carnavaca Rodríguez

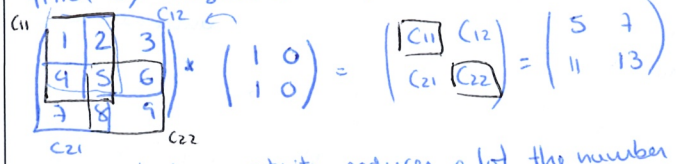
Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.

Backpropagation is an algorithm used for weight optimization of models which follows the oposite direction of the network. This is useful because it uses gradient descent at each step of the network, starting from the output node and go one to the input.



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 instead of acting as weights, act as filters, performing changes in the input by applying a convolution. This convolution is a sum of multiplication of the filter by a region of the input (see example) :



This make us understand how weights can be shared among nodes and also that it has a partial connectivity.

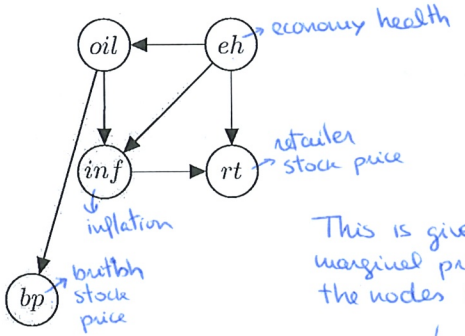
This partial connectivity reduces a lot the number of weights of the model making the training cheaper in terms of computational cost.

In practice they are implemented as a filters (for example for feature extraction from images).

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.

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).



This is given by the combination of marginal probs and conditionals between the nodes and its parent nodes.

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

(2) a) NO, they are dependent. Because we can see that the information flow between "eh" and "bp" is not stopped by any other given variable. So for any given data they are related (head-to-tail)

b) YES, they are independent. Because for a given information of "oil" interrups the unique information flow between "eh" and "bp". This means that it can produce changes on the info, so there is no a direct relation between them (eh and bp). (head-to-tail)

I CONTINUE IN THE SHEET

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 $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: MIRIAM CARAVACA RODRÍGUEZ

Assignatura: MACHINE LEARNING Grup: Curs:

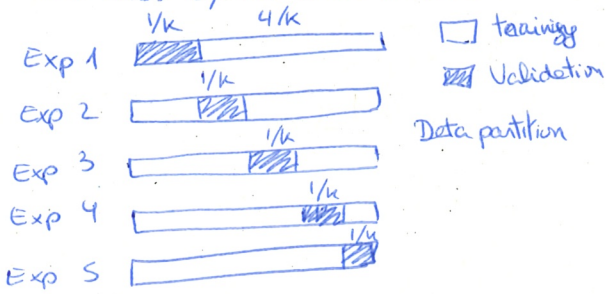
Centre/Estudi:

Professor/a:

Data: 14/12/2021

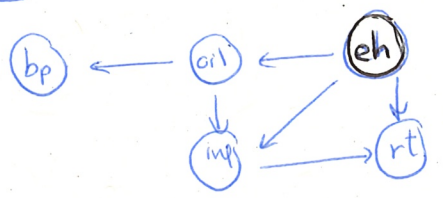
Q.4

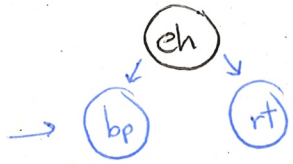
Another thing can be done is to perform a cross-validation. The path is the same one, but for each model we compute the training and validation experiment k times for different spliting of the data. See diagram. Then we compute the mean of the validation loss among the experiments and compare them between models. This gives us a better representation of the true loss by the validation loss.



Q.9

c)





No, they are independent, because we have a (tail-to-tail) node, where both "bp" and "rt" have "eh" as a parent node. This means that changes in "eh" will affect in "bp" and "rt".

Universitat Pompeu Fabra Barcelona

Nom i cognoms: MIRIAM CARRASCO RODRÍGUEZ

Assignatura: MACHINE LEARNING Grup: Curs:

Centre/Estudi:

Professor/a:

Data: 14/12/2021

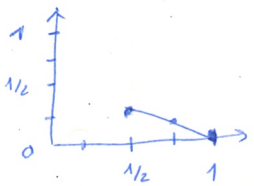
Q.2 $\max xy \rightarrow x+2y=1$ From this relation we can see that for obtaing the max we need that x and $y > 0$. Otherwise we will have a multiplication $(+)\cdot(-) = (-) \rightarrow$ do not achive the max.

We can isolate x from $x+2y=1 \rightarrow x = 1-2y$ and substitute in the max equation

$$\max xy = \max (1-2y)y = \max y - 2y^2$$

To find the max or min of a function we can compute $\frac{d}{dy}(y-2y^2) = 0$

$\frac{d(y-2y^2)}{dy} = 1-4y = 0 \rightarrow \boxed{y^* = \frac{1}{4}}$ to meet the requirement $\boxed{x^* = 1-2\frac{1}{4} = \frac{1}{2}}$



$\begin{cases} y=0 \\ x=1 \end{cases}$ $\begin{cases} y=\frac{1}{8} \\ x=\frac{3}{4} \end{cases}$

From $x = 1-2y \rightarrow$ To make $x>0 \rightarrow y < \frac{1}{2}$

From $y = \frac{x-1}{-2} \rightarrow$ To make $y>0 \rightarrow x<1$

$\downarrow$

$y = \frac{1-x}{2}$

So the optimal values for this problem is $\begin{cases} x = \frac{1}{2} \\ y = \frac{1}{4} \end{cases}$

The maximum value of the function $\max_{x,y} xy$ is $\max (1/2 \cdot 1/4) = \frac{1}{8}$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Sergio Calo Oliveira

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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

$f(x)$ is convex if $\begin{cases} f'(x) \neq 0 \\ f''(x) > 0 \end{cases}$

$f(x) = e^x \rightarrow f'(x) = e^x \;\|\; f''(x) = e^x$

$f''(x) = e^x > 0$ for $x > 0$. q.e.d

$f(x) = x \log x \rightarrow f'(x) = \log(x) + 1 \;\|\; f''(x) = \frac{1}{x}$

$f''(x) = \frac{1}{x} > 0$ for $x > 0$. q.e.d.

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.

$x + 2y = 1 \rightarrow x = 1 - 2y$

$\max xy \rightarrow f(y) = (1 - 2y)y = y - 2y^2$

$f'(y) = 1 - 4y = 0 \rightarrow \boxed{y = \frac{1}{4}} \;\|\; x = 1 - 2 \cdot \frac{1}{4} = \boxed{\frac{1}{2} = x}$

$\max x \cdot y = \frac{1}{4} \cdot \frac{1}{2} = \boxed{\frac{1}{8}}$

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 algorithm is to organite (classify) a set of unlabeled data into k different clusters (classes). this is then an unsupervised learning algorithm.
the algorithm initialize k centroids randomly, and it assign (~~this centroid to the~~) every sample to the closest centroid. Then, it recalculates the centroid by computing the mean of the samples and repeats the process (~~iterativity~~) until it converges.

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?

When we are training a model, we want this model to be able to generalize beyond the train set. If we evaluate the performance of the model using the traing set, we can not be sure that it can have the same performance with unseen data. That's why we keep ~~some~~ part of our data outside the training set. Now, in order to evaluate the performance of a model* (continue)

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
- pros: interpretable outputs
- cons: flexibility

Support vector M.
- pros: efficience; using the kernel trick we can classify non-linear problems.
- cons: problems to handle misslabeled data, only ~~non~~ linear problems if we don't apply kernel trick

Neural networs
- pros: powerful (they can approximate every possible function with enough parameters. Universal aproximation theorem)
- cons: not interpretable (black box)

Name: Sergio Calo Oliveira

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

BP is an optimization algorithm which can iteratibily minimize the loss function by computing its gradient with respect to the weights of each layer. The step size in this descent of the gradient is parametrize by the learning rate (I will call it α)

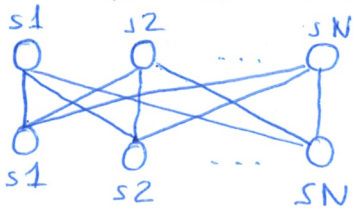
$w^l_{t+1} = w^l_t - \alpha \frac{\partial L}{\partial w^l_t}$, where L is the loss, w the weight matrix, l=layer (continues)

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?

A convolutional layer is ~~made~~ build by a kernel of a fixed size. This kernel moves along the input by jumps of a length parametrized by the stride. At every step, this kernel ~~perform~~ computes the convolution of the inputs inside the kernel with the weights of the kernel, giving a scalar value as the output at each step. This operation is useful in deep (continues)

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 the attention mechanism, given an input sequence, you have a graph connecting every element to all the elements in the segcence:



(continues)

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).

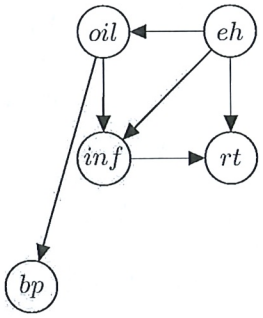


Figure 1: A simple Bayesian network

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?

1. joint probability: ~~p(bp) = p(eh) · p(oil|eh) · p~~
$p(eh, oil, rt...) = p(eh) \cdot p(oil|eh) \cdot p(inf|oil) \cdot p(rt|eh, inf) \cdot$
$p(bp|oil)$

2. a) No, they are dependent. This is because we can find a path between them with a head-to-tail relation.
eh → oil → bp

* (continues)

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 $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

2. The Bayes theorem gives us this relation between θ and D: $P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$
where $p(\theta)$ is our prior. So, if we assume a prior value to θ, we can update this value when we know new information about D using this relation. ~~(Bayes ...)~~

Universitat Pompeu Fabra Barcelona

1

Nom i cognoms: Sergio Calo Oliveira

Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Q 4.* we will ~~use~~ compute the loss over ~~this~~ unseen data, the so called validation ~~loss~~.

We are asuming that this loss is the loss of the model in the whole data distribution (the true loss), so we choose the model with the lowest validation loss.

However, we can notice that our validation set is finite, and it can not represent the whole data distribution of the problem. This validation loss is, then, just an aproximation to the true loss. Usually, when we sample more datapoints ~~from t~~ to the validation set, the aproximation tends to improve.

Q.6* ~~What makes this algorithm very efficient is that~~ You start computing the gadient at the last layer and you continue moving through the other layers backwards. What makes this algorithm very efficient is that at each layer, you only need to compute the gradient for the weight matrix of that especific layer.

Q7* learning because it has the power of act as ~~extract~~ a "filter" to extract more relevant features from the original input.

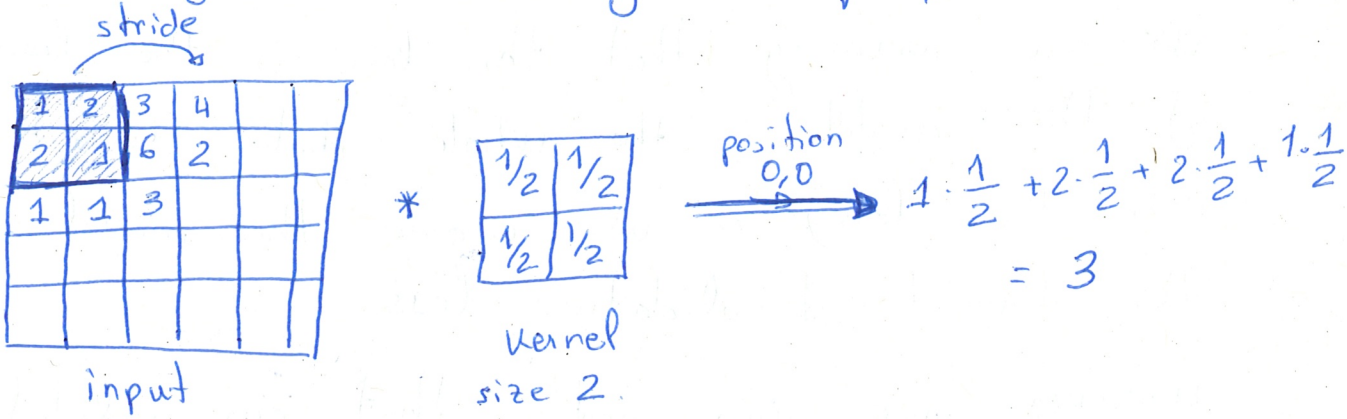


Fig: Example of a 2D convolutional layer performance

However, the ~~weights~~ parameters that build the kernels ~~have to be~~ need to have the appropiate value. To do so, we will learn this value by introducing them in the BP algorithm. Doing so, we need to compute how the loss function depends on the value of this parameters, by computing its gradient.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Sergio Calo Oliveira

Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

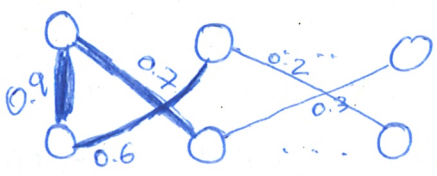
Professor/a:

Data:

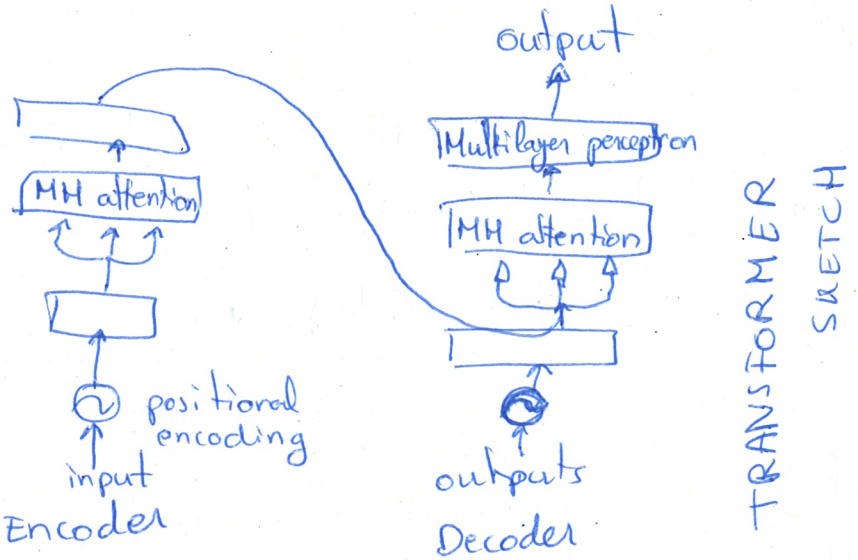
Now, the new kernel parameters are:

$$w^k_{t+1} = w^k_t - \alpha \nabla_{w^k} Loss$$

Q.8* The main idea here is that every node has associated a vector ("key" on one side, "query" on the other side). Now, given this vectors, you can compute the "importance" of the relations between the nodes by computing the scalar product of the key and the query. In this way, if the key and the query have the same direction (or similar), the value of the connection between them (attention) is high.



In the Transformer network, the input sequence and the positional encoding information are passed through the network so this key/query values can be learned (among other things). this $ attention mechanism lives in the multihead attention layers, where this operation happens. This attention relations are very useful $ in order to make the right predictions since we can focus on those features that are more relevant for each case.



Q. 9*

b) If the value of "oil" is given, the previous path is broken. Other path between "bp" and "eh", following "inf":

Is also broken when we know "oil". Then, bp and eh are conditionally independent when

Nom i cognoms: Sergio Calo Oliveira

Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

Professor/a:

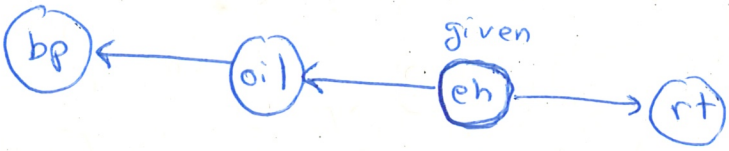
Data:

we know oil.

c) The path between rt and bp can be:

Path A.

or



Path B.

The path B is broken when "eh" is given. However, Path A is still open. Therefore, rt and bp are dependent given "eh".

Machine Learning 2021-22
Final Exam

14 December 2021

Name: ANDREU PASCUET FONTANET

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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

Defining a function $f(x)$ is convex if $f''(x) > 0$. Therefore for $a(x) = e^x$:

$a''(x) = e^x$, if $x > 0$ then $e^x > 0$, so $a''(x) > 0$ meaning $a(x) = e^x$ is a convex function.

Similarly, for $b(x) = x \log x$:

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

$b''(x) = \frac{1}{x}$, if $x > 0$, then $\frac{1}{x} > 0$, so $b''(x) > 0$ meaning $b(x) = x \log x$ is also 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.

SOLVED IN AN EXTRA SHEET

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 technique that seeks to group a set of data into different subgroups, minimizing the difference between the points of a subgroup and its centroids.

For that, the K-means algorithm follows these steps:

1. Randomly initialize K centroids.
2. Assign datapoints to the closer centroid (calculate the difference between each datapoint and centroid, and select the centroid with minimum difference).
3. Reposition the centroids to the mean of its assigned datapoint.
4. Repeat from point 2 until convergence of the centroids is meet.

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?

From the training data for a model, a percentage of this data can be kept and not used in the training, to later on be used as a validation set. This way, we can train and validate different models for which, at the end, we will choose the one with a better validation performance and train with the whole training set.

This process has the advantage of better estimating the true loss because it is using unseen data by the model, while other process, only using the empirical loss only take into consideration already seen data, making the model more prone to overfitting.

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
  - Advantage: easy to interpret
  - Disadvantage: prone to overfit (large number of nodes)
- Support vector machines
  - Advantage: highly dependent on C
  - Disadvantage: map non-linear information to linear in higher dimensions
- Neural networks
  - Advantage: feature decision can be automated
  - Disadvantage: difficult to interpret

Name: .....

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

Backpropagation algorithm in supervised learning is based on approximating the calculated weights by minimizing a loss function that considers the difference between the real label and the predicted value given an input. Therefore, we first need to calculate the output and then go backwards to update the weights, that is why it is called backpropagation.

Backpropagation uses the gradient descent when no analytical solution for updating the weights is available e.g.: $w_{t+1} \leftarrow w_t - \eta \nabla_w L$, where η is the step size. For backpropagation to be efficiently implemented these η must be as large as possible to let the model converge fast, but small enough so that it does not suppose divergence. Therefore, it is possible to start with large η and gradually reduce them as the model converges. It is also possible to apply stochastic gradient descent instead of gradient descent if a huge training set is provided. In... continues in extra sheet

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 a series of filters over the data that reduce or amplify the number of features depending on the filter size. The great advantage of convolutional layers is that they do not need to be fully connected, as well as the number of weights does not depend on the number of nodes, but on the number of filters and its size.

Convolutional layers are usually implemented in hand of maxpooling layers, which reduce feature dimensionality by selecting predominant features, making the model more efficient and reducing the propagation of possible errors (such as those incorporated... continues in extra sheet

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.

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).

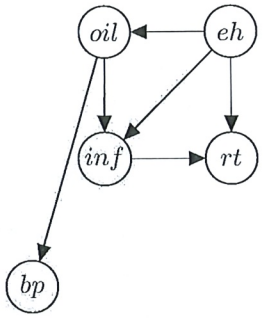


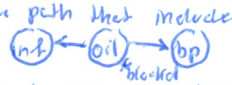
Figure 1: A simple Bayesian network

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?

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

(2) a eh ⫫ bp because we have a head-to-tail node that is not blocked

b eh ⫫ bp | oil because the above head-to-tail node now is blocked and, moreover, the other possible path that includes a tail-to-tail node is also blocked (all paths are blocked)



c rt ⫫ bp | eh because we observe an open path with a head-to-tail and a tail-to-tail node:

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) The likelihood function for this model would be $p(D \mid \theta)$

(2) Bayes theorem tells us that:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(D)}{\sum_D p(D \mid \theta)\, p(D)}, \quad \text{which is understood as posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Therefore knowing the likelihood and assuming a prior for the model the posterior (parameters θ in light of D) could be calculated.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: ANDREU PASCUET FONTANET

Assignatura: MACHINE LEARNING Grup: Curs:

Centre/Estudi: CBME

Professor/a:

Data: 14 / 12 / 2021

~~QUESTION 2 (CONTINUATION)~~

~~Above we can minimize by setting the gradient to 0:~~

~~$\nabla_{x,y,\lambda} \mathcal{L}(x,y,\lambda) \in 0 \Leftrightarrow$~~

~~So we have obtained the optimal value of x and y depending on λ. For determining the~~
~~value of λ which maximizes we:~~

~~$\nabla_{x,y} \mathcal{L}(x^*, y^*, \lambda) \in 0$~~

~~$\nabla_{x,y} \mathcal{L}(x^*, y^*, \lambda) =$~~

QUESTION 2

Given the optimization problem we can determine the Lagrangian:

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

and, therefore, its gradient:

$$\nabla_{x,y} \mathcal{L}(x, y, \lambda) = \begin{bmatrix} \frac{\partial \mathcal{L}(x,y,\lambda)}{\partial x} \\ \frac{\partial \mathcal{L}(x,y,\lambda)}{\partial y} \end{bmatrix} = \begin{bmatrix} y - \lambda \\ x - 2\lambda \end{bmatrix}$$

↳ CONTINUES BACKWARDS

↳ CONTINUATION QUESTION 2

for optimizing we set the gradient to zero to find optimal x and y:

$$\nabla_{x,y} \mathcal{L}(x,y,\lambda) = 0 \rightarrow \begin{bmatrix} y-\lambda \\ x-2\lambda \end{bmatrix} = 0 \Rightarrow \begin{matrix} y^* = \lambda \\ x^* = 2\lambda \end{matrix}$$

To get rid of the dependency on λ we can compute the dual of the Lagrangian, that is:

$$\begin{aligned} \mathcal{L}(x^*, y^*, \lambda) &= 2\lambda \cdot \lambda + \lambda(\lambda - 2\lambda - 2\lambda) = \\ &= 2\lambda^2 + \lambda(\lambda - 4\lambda) = \\ &= 2\lambda^2 + \lambda - 4\lambda^2 = \\ &= \lambda - 2\lambda^2 \end{aligned}$$

and setting the gradient to 0:

$$\nabla_\lambda \mathcal{L}(x^*, y^*, \lambda) = 1 - 4\lambda = 0 \rightarrow \lambda = \frac{1}{4}$$

Therefore, substituting at the x^*, y^* variables:

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

$$y^* = \frac{1}{4}$$

and substituting at the objective we get its optimal value:

$$\max_{x,y} (xy) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \quad //$$

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

Universitat Pompeu Fabra Barcelona

Nom i cognoms: ANDREU PASCUET FONTANET

Assignatura: MACHINE LEARNING Grup: Curs:

Centre/Estudi: CBME

Professor/a:

Data: 14 / 12 / 2021

QUESTION 6 (CONTINUATION)

... this case, the gradient is not calculated over the whole dataset, but only a portion. Which in expectancy should have the same value as the global. In terms of efficiency the stochastic gradient descent needs to define a batch size which should be as large as possible without compromising the computational effort.

QUESTION 7

... by padding to avoid shrinking). Moreover, final layers of a convolutional neural network use to be feedforward layers, as in those cases data dimensionality has been reduced by the convolutional layers and they are better at inferring high-holdility features.

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Julio Cesar Casas Quiroz

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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

1° $\frac{\partial}{\partial x} e^x = e^x \rightarrow$ 2° $\frac{\partial}{\partial x} e^x = e^x$, $x > 0 \rightarrow e^x > 0 \rightarrow e^x$ convex for $x > 0$

1° $\frac{\partial}{\partial x} x \log x = x' \log x + x \log(x)' = 1(\log x) + x\left(\frac{1}{x}\right) = \log(x) + 1$

2° $\frac{\partial}{\partial x} \log(x) + 1 = \log(x)' + 1' = \frac{1}{x} + 1$, $x > 0 \rightarrow \frac{1}{x} + 1 > 0 \rightarrow x \log x$ convex for $x > 0$

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.

- $x + 2y = 1 \rightarrow y = \frac{1 - x}{2}$
- $\frac{x(1-x)}{2} = \frac{x - x^2}{2} = \frac{x}{2} - \frac{x^2}{2}$

Derivate
- $\left(\frac{x}{2} - \frac{x^2}{2}\right)' = \left[\frac{1}{2}x^0 - \frac{2x^1}{2}\right] = \frac{1}{2} - x$

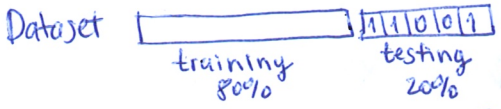
optimal
- $\frac{1}{2} - x = 0 \rightarrow x = \frac{1}{2}$, $y = \frac{1 - 0.5}{2} = \frac{0.5}{2} = 0.25$
- $x = 0.5$, $y = 0.25$, $xy = \frac{50^1}{100_4} \times \frac{25^1}{100_2} = \frac{1}{8} = 0.125$

Julio Cesar Casas Quiroz

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 is to cluster the data in similar groups, but~~
Unsupervised centroid-based algorithm that groups data ~~in clusters~~ with similar characteristics, but each cluster is different between them. 1) Randomly initialize centroids, 2) assign each point of data to the most closest centroid (by euclidean distance i.e.), 3) recalculate the centroids of each cluster, by the center of the points of each cluster, 4) Repeat 2 and 3 until neither point changes of cluster

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?



- In the process of validation, we take a sample of the training dataset (ie. 20%) and exclude it from the training process.
- Then the h(x) learned is applied in the testing sample, and we can compare the prediction and the testing label.

h(x) = 0 1 0 0 1
x ✓ ✓ ✓ ✓
accuracy = 0.8

- We can perform many training processes with different algorithms and hyperparameters and select the one with least ~~error~~ validation loss.
- Validation loss is ~~the~~ an optimistic estimate of true loss because it's only a sample of what happens in the past. So we don't know the true in the future for validation.

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
Decision Trees	• Good explainability because of rules	• Usually high bias for complex problems
SVM	• Usage of non-linear kernels to solve particular problems (usage of Gaussian kernel)	• ~~Could~~ Could be very expensive in Hardware resources ~~computation~~ computation for high dimensionality problems.
NN	• Good applications with unstructured data as images and text and audio	• Challenges in overfitting because of its complexity.

Name: Julio César Casas Quiroz

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

The backpropagation algorithm is used to recalculate the weights to get the optimal ones. First, we have to obtain all the gradients of each node, by derivating the loss function respective of each node. Then, we calculate the gradient descent to get the optimal weights. To solve it efficiently is recommended to use matrix computation which can be solved fine with python or matlab.

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 helps to reduce dimensionality in CNN problems (i.e. images, could be so much expensive to compute with FCN). But, keeping a good representation of the input and reducing noise.
First, input data is convolusional by passing through a matrix with initialized weights and selected shape, then an activation function is applied (i.e. signoid), and we have a max pooling matrix, that finally will be flatten to do the final prediction.
The change from backpropagation standard, is that we have to find the optimal weights of the convolution, and now we have a matrix of weights to optimize for each convolution.

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.

To include in the analysis not only the past events in a sequence, but also analyze the next events.

Name: Julio Cesar Casas Quiroz

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).

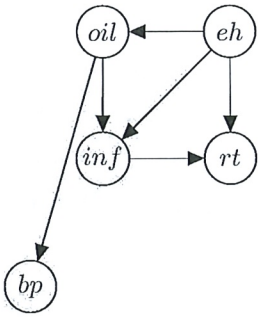


Figure 1: A simple Bayesian network

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 eh-oil-bp is not blocked

b) yes, because oil node is blocking, and all paths goes through oil.

c) no, because rt-eh-oil-bp is blocked by eh, but rt-inf-oil-bp is not blocked

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) Likelihood $= P(D|\theta) = \prod_x P(x|\theta), x \in D$

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

Machine Learning 2021-22
Final Exam

14 December 2021

Name: STEPHANIE RODRIGUEZ OSORIO

- **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 $\frac{\partial}{\partial x} (f(x) \cdot g(x)) = \frac{\partial f(x)}{\partial x} \cdot g(x) + f(x) \cdot \frac{\partial g(x)}{\partial x}$.

$e^x \times \log x$

$\frac{d}{dx} e^x = e^x$

$\frac{d}{dx} \log x = \frac{1}{x}$

$\frac{d}{dx}(f(x) \cdot g(x)) = \frac{df(x)}{dx} \cdot g(x) + f(x) \frac{\delta g}{\partial x}$

$\frac{d}{dx} x^p = p \cdot x^{p-1}$

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.

$\max_{x,y} xy$ $\quad x + 2y = 1$

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 K-MEANS clustering objective is to create a ~~specific~~ group of ~~means~~ data. near the solution. ~~es~~pected ~~related~~ grouping them for their neighborhood similarity.

* It Approximate to the solution by using the values. recogniced as the k-means and gererated a media of them to use it as the aproximation solution.

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?

* for usign validation. we have to..

* Is because the validation loss. is not realy the true loss and only helps you to. have a estimated near to the one that represent the true loss. in a general way.

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.

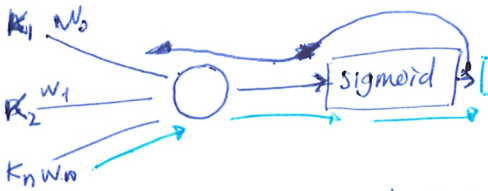
DECISION TREES → (A) ~~easy to use~~ basic supervised learning algorithm.
(D) is not so accurate than the other ~~two~~

SVM. → (A) it let generalize the data, (flexible) ~~easy~~ using a kernel depending on the application.
(D) the ~~computational~~ computational cost

NN → (A) it works well for image processing Applications.
(D) Requires a big amount of data to train it ~~the~~ ~~to~~,

Name: STEPHANIE RODRIGUEZ OSORIO

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



* THE Algonthm works like, the first time, OR At it start it assigne as he wants the weights of the values (data) that enters to the nn and it depends on this the value of the outside first training, then with this the nn calibrate the new weight of the data enter and do this until there is a good set of weights for the back propagation (show in the image).

* a efficient implementation of it is to close the attention on each ~~ejaration~~ of the nn, at least in the first couple of them, be sure it is distributing the weights apropiately

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 CONVOLUTIONAL LAYERS ~~are~~ ~~works~~ ~~in the~~ creating a ~~key~~ specific section of the layers in which the data have to convert to obtain the specific outcome

*

It have to be clear

* ~~in~~ in which way it wanted to converge the backpropagation so ~~that~~ it will perform the convolution ~~in the~~ ~~backpropagation~~ in the section of the specific layer of the backpropagation algorithm

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~~ called attention is a way to focus on the ~~other~~ information that we want in order to work with it as the importance the evaluation of the deep learning do in the learning..

Name: STEPHANIE RODRIGUEZ OSORIO

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).

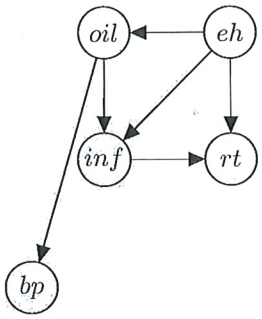
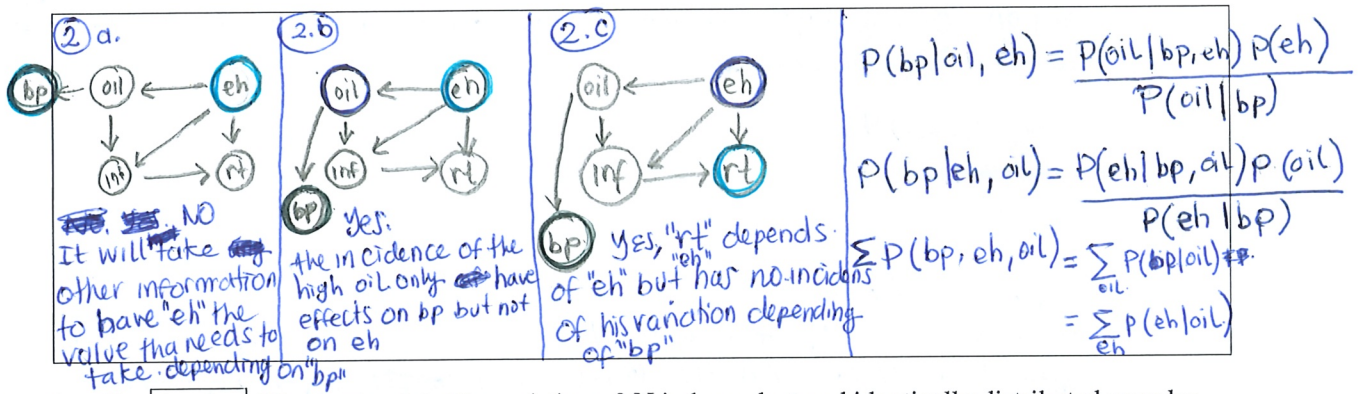


Figure 1: A simple Bayesian network

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?



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 Likelihood → $\binom{D}{N}\binom{N}{L}\binom{D}{\sigma^2}$

② BAyes theorema can be used to update the parameters θ by changes the values in terms of the probability of the events happening, focusing on being evaluated by D.

upf. Universitat Pompeu Fabra Barcelona

Nom i cognoms: Huicheng Zhang

Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Q1: $\because \frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x}(\frac{\partial}{\partial x} e^x) = \frac{\partial}{\partial x} e^x = e^x$. $e^x > 0$ for $x > 0$

$\therefore$ univariate function e^x is convex

$\because \frac{\partial}{\partial x}(x \log x) = \frac{\partial}{\partial x}(x) \cdot \log x + \frac{\partial}{\partial x}(\log x) \cdot x = \log x + 1$

$\frac{\partial}{\partial x}(\log x + 1) = \frac{1}{x} > 0$ for $x > 0$

$\therefore$ Univariate function $x \cdot \log x$ is convex

Q2: $\because$ S.t. $x + 2y = 1$

$\therefore x = 1 - 2y$

max xy can be transformed into an optimization problem of a univariate function

$$\max_{x,y} xy \quad \max_{y} (1-2y)y$$

$\because$ suppose $f(y) = (1-2y)y$

$\therefore f'(y) = -2 \cdot y + (1-2y) = 1 - 4y$

$f''(y) = -4 < 0$,

$\therefore$ when $1 - 4y = 0$, $f(y)$ reach maximum

$\therefore f(y)_{max} = f(\frac{1}{4}) = \frac{1}{8}$

$y = \frac{1}{4}$, $x = \frac{1}{2}$

Q3: The objective of k-means clustering is to find an optimal centroid for each cluster. The algorithm is the following:

① Initialize the centroid in each cluster

② Compute the loss function which measures the weighted sum of distances of all samples in one cluster to the centroid $L_s(C_k) = \frac{1}{m} \sum_{i=1}^{m} \|x_i - C\|$ $C_k \in (C_1 - C_k)$

③ Update centroid

④ Compute new cluster

⑤ Repeat ②-④ until $L_s < \varepsilon$

Q4: The process:
① ~~Split~~ Split the training set D into validation set V and training set D'. V has an independent distribution of D.
② Suppose the hypothesis set is $H=\{H_1, \dots H_m\}$, we train each $H_i \in H$ with D' and calculate the loss function in the validation set V.
③ Select the hypothesis with minimum validation Loss L_V, train the model with the original training set D.
Reason why the validation loss an optimistic ~~estic~~ estimate of the true loss:
Because L_V is less likely to be overfitting than training loss. Validation set is unseen during training. K-fold validation is useful ~~is~~ in practice.

Q5:
Decision Trees: It is interpretable and allows expert knowledge to be added in the model; However, it can't deal with real-value feature because it is hard to split the node base on realvalue feature.

Support vector machine: The design of margin and support vector makes it robust to noise in data and ~~perform well in linearly separable data~~ the classification result is close to true labeling function. However, SVM only works well in ~~the~~ linearly separable data originally. The design of large margin can lead to overfitting

Neural Network: Strong ability to approximate the true labeling function.
High computation cost since the number of weights are high.

Q6: Backpropagation algorithm is used to update the weights w within the hypothesis. When the loss function is convex, smooth and bounded, ~~der~~ derivative will be computed from the output layer to input layer so that all the weights would be updated.

Backpropagation can be efficiently implemented by constructing a matrix of calculating derivatives. Matrix mulplication is fast, weights would be updated as ~~w = w - α ∂x~~

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Huicheng Zhang

Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Q7: In deep learning, convolutional ~~layer~~ manipulation applied element-wised mulplication to the data within the kernel's receptive field, and output one value (which is the sum of mulplication) to the convolutional layer. Weights are shared within convolutional layer. The design of kernel focus on the locality of the input vector.

In practice, a new matrix which reorder the input vector according to the size of kernel are implemented, in order to utilize the fast matrix computation speed. Padding is implemented to the input layer in order to increase the dimension of convolutional layer. Stride can be manipulated to decrease the dimension of convolutional layer.

Because of the weight sharing nature of convolutional layer, weights ~~are~~ got updated simultaneusly when doing standard backpropagation. All the weights in the convolutional layer are the same.

Q8: Attention includes Long-short term memory (LSTM) unit and ReLU gated layer in order to learn about the context of the input and assign different weights to the sequence (focus on different components of the input)

Q9: (a) eh is not independent of bp, because eh - oil - bp is a subset without any subset blocks in between. Information can flow from eh to bp.

(b) eh ⊥⊥ bp | oil = high. because the ~~obs~~ observed oil blocks eh and bp, Information couldn't flow from eh to bp. Although ~~eh~~ information can flow from eh to inf, since bp - oil - inf is tail to tail ~~relat~~ structure for oil, information couldn't pass oil to bp. Thus eh is independent of bp if oil is observed

(c) rt is not independent from bp, because the path rt - inf - oil - bp carries the information without any blocks in between.

(d) $P(bp) = P(bp|oil)\,P(oil) = P(bp|oil)\,P(oil|eh)\,P(eh)$

$P(oil) = P(oil|eh)\,P(eh)$

$P(eh) = P(eh)$

$P(inf) = \cancel{P(inf|oil)\,P(oil)\cdot P}.$

$P(inf|oil, eh)\,P(oil, eh)$

$=$

Q10. (1) Bayes: $P(\theta|D) = \dfrac{P(D|\theta)\,P(\theta)}{P(D)}$

∴ The likelihood function of θ for fixed D is $P(D|\theta)$

(2) Parameter θ can be updated by ~~maxima~~ maximizing $P(D|\theta)\,P(\theta)$,

$\theta = \arg\max_{\theta} P(D|\theta)\,P(\theta) = \arg\max_{\theta} P(D|\theta)\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-u}{\sigma})^2}$ ~~Since~~

Compute the gradient of $\nabla \arg\max_{\theta} P(D|\theta)\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-u}{\sigma})^2} = [\frac{1}{2}\frac{2\theta}{\sigma^2} - \frac{1}{2}(\frac{\theta-u}{\sigma})^2]\, e^{-\frac{1}{2}(\frac{\theta-u}{\sigma})^2} * P(D|\theta)$

get the θ that can maximize the equation above ∴ $\theta =$

since data D consists of N independent and identically distributed samples

∴ $\theta = \arg\max_{\theta} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-u}{\sigma})^2}$ ~~*~~ =