Machine Learning 2021-22
Final Exam

14 December 2021

Name: EMMANOUIL P. ALAIDOGOS

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

- If $f''(x) \geq 0$ (for f: univariate) it follows that: $\left(\frac{f'(y) - f'(x)}{y - x} \geq 0, \forall x, y \right)$
- For $f(x) = e^x \Rightarrow f'(x) = e^x \Rightarrow f''(x) = e^x > 0 \ \forall x > 0$.
- For $f(x) = x \log x \Rightarrow f'(x) = \log x + x \frac{1}{x} = \log x + 1 > 0$

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

$$\max_{x,y} xy$$
$$\text{s.t. } x + 2y = 1 \Rightarrow \underbrace{1 - (x + 2y) = 0}_{g(x,y)}.$$

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

Lagrange: $\nabla f = \lambda \cdot \nabla g$

- $\mathcal{L}(x, y, \lambda) = xy - \lambda [1 - (x + 2y)]$
- K.K.T.: $\nabla \mathcal{L}(x, y, \lambda) = 0 \Rightarrow \begin{cases} \frac{\partial \mathcal{L}}{\partial x} = y + \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial y} = x + 2\lambda = 0 \end{cases} \Rightarrow \boxed{\begin{array}{l} y^* = -\lambda \\ x^* = -2\lambda \end{array}}$ (I)

(dual) and

- $\mathcal{L}(x^*, y^*, \lambda) = 2\lambda^2 - \lambda - 4\lambda^2 = -\lambda - 2\lambda^2$

Maximize dual: $\nabla_\lambda \mathcal{L} = 0 \Rightarrow -1 - 4\lambda = 0 \Rightarrow 4\lambda = -1 \Rightarrow \boxed{\lambda = -\frac{1}{4}}$ (II)

So: (I) $\overset{\text{(II)}}{\Rightarrow}$ $\begin{cases} y^* = \frac{1}{4} \\ x^* = \frac{1}{2} \end{cases}$

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

- In k-means algorithm, we try to minimize the cost:

$$G_{k\text{-means}}(C_1, C_2, \dots, C_k) = \sum_{i}^{k} \sum_{x \in C_i} d(x - \mu(C_i))^2$$

where $\{C_1, C_2, \dots C_k\} \in S$: partition of dataset.
and $d(\,,\,)$ the distance metric (eg. distortion $\sum_i r_i (x_i - y_i)^2$).

- In every step the $G_{k\text{-means}}$ is monotonically descending.
→ Initially we choose random $\mu(C_i)$ centroids.
→ We create C_i' clusters based on μ_i
→ We update centroids so that $d(x_i - \mu'(C_i'))^2 \leq d(x_i - \mu(C_i))^2$.
→ Repeat till centroids do not change.

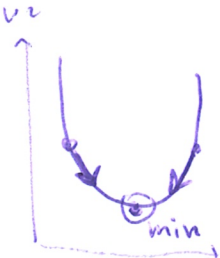
Question 4: [1 point] Describe the process of using validation to perform model selection. Why is the validation loss an optimistic estimate of the true loss?

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.

	PROS ⊕	CONS ⊖
Decision trees	No need to store the data (we just 'save' the distributions based on the conditions)	There is no guarantee we get a global minimum. It's a greedy algorithm (seperates for maximum IG and it stops to first result of pure leaves).
Support vector mach.	Using quadratic optimization, really easy to converge and if we combine with $K_{ij} = \langle \phi, \phi \rangle$ kernels it can become really powerfull on seperating high-dimens. data.	Many parameters to optimize (C, γ ...)
Neural Nets	Can solve complex problems that need more than one hyperplanes.	They require a big amount of data in order to train and they can easily overfit.

Name: EMMANOUIL PALAIOLOGOS

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



The goal on backprop. algorithm is to gradually descend in the weight space, towards the direction of $-\nabla L$ (opposite of gradient of loss) till we reach the minimum. We take steps, dictated by a learning rate η so that $w \leftarrow w - \eta \nabla L$ (being carefull not to choose extreme values of η that can prevent the algorithm from convergence). This process though can become really slow, so we prefer work on subsets of the data s'CS, with "partial" gradients (eg. batching).

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?

CNN's are so powerful, due to the weight-sharing property between each layer. They are based on convolution operations on an initial pixel grid with a "filter". Convolving leads to a new "volume", with length and height that depend on the size of filter and the amount of "slide"s. The depth of this new "volume" is given by the number of filters applied. The idea is that the algorithm will learn this filters. On each layer the filters are able to detect more complex pattern, as we go "deeper". We usually apply a "max" or "average" pooling operation after each conv. layer, and we ~~flatten all "volumes" to a single dimensional array to feed a fully~~ connected network.

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: EMMANOUIL PALAIOLOGOS

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).

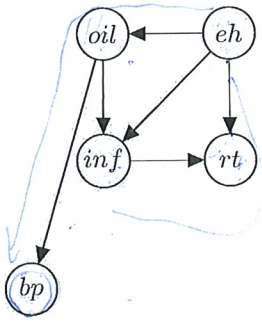


Figure 1: A simple Bayesian network

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

(2) (a) Path between "bp" and "eh" is open, hence "eh" is dependent of "bp" (head-to-tail)

(b) Path "eh" – "oil" – "bp" blocked on "oil" and since we have a head-to-tail, "eh" and "bp" become independent.

(c) Path is blocked on "eh", (not a tail-to-tail connection) so "rt" – "oil" become (independent) and hence "bp" and "rt" as well!

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 of θ to get a specific $\mathcal{D}$: $p(\mathcal{D} \mid \theta) = \prod_i^N p(x_i \mid \theta)$ (assuming independence)
or in general $\prod_i p(y_i \mid x_i, \theta)$ (y_i: targets)

(2) We can maximize the aposteriori $\propto p(\mathcal{D} \mid \theta) \cdot p(\theta)$, where $p(\theta)$ is the prior which also follows the Gaussian ($p(\theta \mid \mathcal{D})$).
Considering the $\log p(\mathcal{D} \mid \theta)$ will make the calculations more feasible for the computer. Every $\log \prod_i p(\cdot)$ becomes $\sum_i \log p(\cdot)$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Miguel Silva Fuentes

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

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

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

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

$\frac{\partial x \log x}{\partial x} > 0$ ∴ so it is convex

$\frac{\partial e^x}{\partial x} = e^x$ $\quad e > 0$, $x \to 0$

$\frac{\partial e^x}{\partial x} = e^x > 0$ ∴ it is convex

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.

$f(x) = \max_{x,y} xy$

$g(x) > x + 2y = 1$

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

K-means: Is a unsupervised algorithm, that help us to classify the data according to a number of generated centroids, the objective is to minimize the distance between every point and the nearest centroid. How it works? First generate $k(N=K)$ random points and it use these points as centroids, the calculate the distance between the nearest points and the centroid. And tries to minimize it, the problem is that sometimes get stuck on local minima.

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?

Model selection: Train multiple models with different parameters using dataset (X) (training), with some distribution, if you want to test the different kind of models you use another set of data (validation), with same distribution, so in this point we assume that our test data will have the same distribution as validation, because we use the validation data to pick the best model for our case. So we trust in:

also we don't validate on train because can overfit

validation data distribution $\Rightarrow$ test data distribution $\Rightarrow$ all possible data distribution

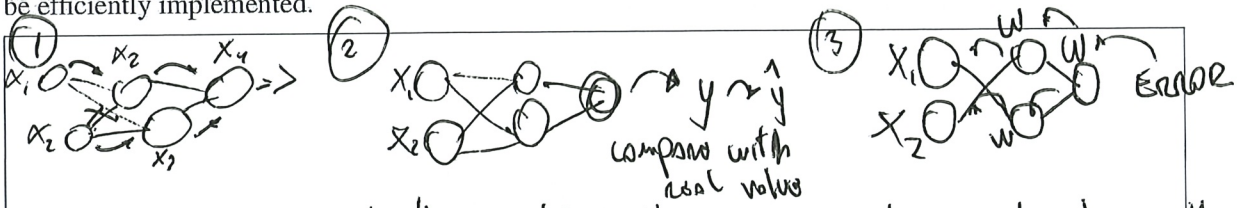
validation loss $\Rightarrow$ test loss $\Rightarrow$ true loss

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.

	Pros	Cons
Decision Trees	- Non linear solver - Fast to train - Can rank variables depending of its. - Can be very flexible	- Prone to overfit - Stuck on local minima - Not always the best split. (depending on implementation)
Support Vector M..	- Can solve linear or non-linear problems - Have regularization inside (minimize the weights automatically)	- Alot of hiperparameters - Has to chose the correct kernel - Slower with more data
Neural Network	- No feature engineering - Improves with data - SOTA in multiple tasks	- Need alot of data - Prone to overfit - Have to chose the right architecture - Black box.

Name: Miguel Silva

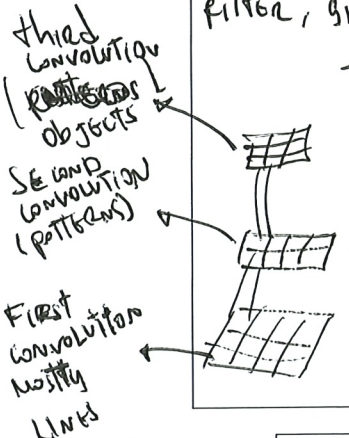
Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.



Backpropagation: correct the weight of the neural network based on the error, basically, go backwards starting from the last layer and change the weight for every neuron if the result is incorrect

Implementation: For the implementation of the backpropagation is basically the chain rule, but we decide how much change ours weights based on SGD.

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?



CNN usually captures the features of the data (images) within the filter, giving with every layer bigger level of abstraction

- to implement this in practice we add layers of filters layers of pooling and sometimes padding to the borders of the image, also pooling can be seen as shrinking
- When we back propagate we only change the values of the filters (the weight of the filters) also we could say that the CNN layer shares 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.

Attention is a novel idea that allows us to "highlight" part of the text (in this case, but can be apply to another kind of data) also merged with a transformer allow the transformer to do parallel training with the same sentence but different attention focus.

Attention make the neural network not to focus on the whole sentence but only on the words that have more meaning inside the text.

Name: Miguel Silva

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).

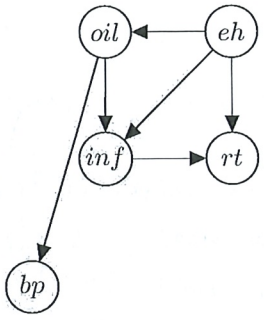


Figure 1: A simple Bayesian network

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) $P(oil, bp, INF, eh, rt) = p(INF|oil, eh)\, p(rt|INF, eh)\, p(bp|oil)\, p(oil|eh)\, p(eh)$
2) a) No, there is a flow of information eh → oil → bp
   b) Yes, because all the ways are blocked eh → oil → bp, head to tail
   c) No, there is a way 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 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) We can use bayes as "inference to model"

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

$P(\theta)$ → PRIOR

$P(D,H)$ → EVIDENCE

$P(D|\theta,H)$ → LIKELIHOOD

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Antonio J. Segura García

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

They will be convex if at minimun point (x_0) we have $f''(x_0) > 0$

We proof that $f''(x) > 0 \quad \forall x > 0$ then if x_0 is in domain these functions will be convex:

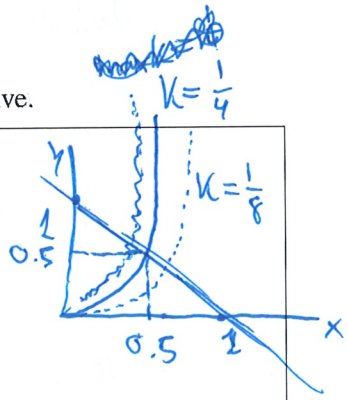
$\frac{\partial}{\partial x} e^x = e^x$; $\frac{\partial^2}{\partial x^2} e^x = e^x \Rightarrow \forall x > 0 \quad e^x > 0 \Rightarrow$ convexe

$\frac{\partial}{\partial x} \log x \cdot x = \log x + 1$; $\frac{\partial^2}{\partial x^2} x \log x = \frac{\partial}{\partial x}(\log x + 1) = \frac{1}{x} \Rightarrow \forall x > 0 \quad \frac{1}{x} > 0 \Rightarrow$ convexe

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.



- Value of the objective $K \Rightarrow x \cdot y = K \Rightarrow y = \frac{K}{x}$
- Constraint: $x + 2y = 1 \Rightarrow y = \frac{1}{2} - \frac{x}{2}$
- If x and y are not negative they must be lower than 1.

x	y	xy
1	0	0
0	1	0
0.5	0.5	1/4
1/4	3/4	3/16 < 1/4
3/4	1/8	3/32

$x = \frac{1}{2} \quad y = \frac{1}{2} \quad K = \frac{1}{4}$

CONTINUES pag 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?

K-means is a unsupervised classification algorithm. It will classifies as a same class the points that are near ~~between~~. It starts from a given centroids μ_i, it compute distance between points and centroids given a metric and it moves centroids in order to minimize ~~the~~ the sum of centroids and points 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?

Given a set of training data, it can be divided in 3 subsets trainning (the biggest), test and validation. Test subset will return the trainning loss or empirical risk and validation can be applied to choose between models or tune the hyperparameters. It is an optimistic estimation of the true loss because it assumes that it has the same distribution than the domain. It reduces bias but increases variance.

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
• DT:	They can be interpretable easier	They dosn't adapt to many problems
• SVM:	efficient linear classificator	Need modification to classify non linear data
• NN:	They adapt to many problems	They can overfitting more than the others

Name: Antonio J. Segura García

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

Neuronal networks can update weights in a very efficient way applying backpropagation algorithm. This algorithm allows to detect which node have to be updated and how. We evaluate a point in NN, if doesn't fit to target result. We will update nodes that activate this response and we will update nodes that activate these nodes iterativelly.

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 connected by a Kernel. In 2 dimensions CNN, Kernel is a NxN matrix and each layer is a bigger MxM matrix (M>N). Kernel stores the weights that will be updated. Input layer elemts are multiplied by Kernel given the value for the next layer then relative position of Kernel and layer is shifted and the process is repeated generating next value.

(CONTINUE page 1)

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.

- One of the difficulties of LSTM in text recognition is link two separated parts of text. Transformers are not feed in a sequence then they can find relationships between two distant words. Each word will show a relationship with others word, these links between words is called "attention".

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).

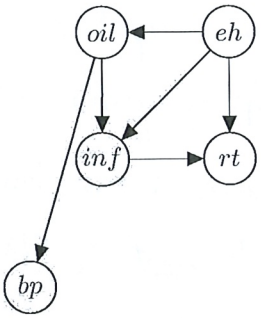


Figure 1: A simple Bayesian network

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

② a) $eh \not\perp\!\!\!\perp bp$ because they are linked by oil node

b) $eh \perp\!\!\!\perp bp \mid oil$ because ~~oil is the node that connects~~ only oil connects eh and bp

c) $rt \not\perp\!\!\!\perp bp \mid eh$ because there is a path 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 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 $p(\theta|D) = \dfrac{p(D|\theta)\, p(\theta)}{p(D)}$

② Bayes theorem can be used to calculate the inverse probability. Then we can update θ and get a bigger $p(\theta|D)$ value.

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Antonio J. Segura García

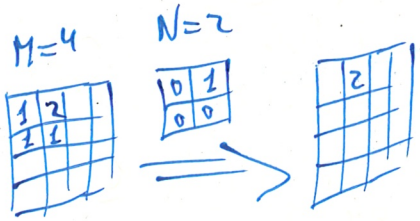
Assignatura: Machine Learning Grup: Curs:

Centre/Estudi:

Professor/a:

Data: 14 December 2021

(7)



• In practice convolutional layer as stored in architectures that have the same dimension than inputs and target, hidden layers can be greater or smaller than input.

• The properties of convolution regarding derivation allows to adapt backpropagation.

(2)

• A better proof: We have to find a relative maximum of

$$x\left(\frac{1}{2}-\frac{x}{2}\right) \Rightarrow \frac{\partial}{\partial x}\left[x\left(\frac{1}{2}-\frac{x}{2}\right)\right]=0 \Rightarrow$$

$$\frac{1}{2}-x=0 \Rightarrow \boxed{x=\frac{1}{2}} \Rightarrow y=\frac{1}{2}-\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{2} \Rightarrow \boxed{y=\frac{1}{2}}$$

$$K=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4} \Rightarrow \boxed{K=\frac{1}{4}}$$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Kosuke James Nishida

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 $f(x) = e^x$,
$f'(x) = e^x$
$f''(x) = e^x$,
therefore, the second derivative $f''(x)$ is positive for $x > 0$, satisfying condition for convexity

for $f(x) = x\log(x)$
$f'(x) = \log(x) + x(\frac{1}{x}) = \log(x) + 1$
$f''(x) = \frac{1}{x}$.
Therefore, for all $x > 0$ the value for $f''(x)$ is positive. Thus $x\log(x)$ 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.

$\mathcal{L}(x, y, \lambda) = xy + \lambda(x + 2y - 1)$.
$\nabla_x \mathcal{L} = 0 = y + \lambda$
$\nabla_y \mathcal{L} = 0 = x + 2\lambda$
$\therefore x^* = (y + \lambda)$, $y^* = (x + 2\lambda)$.
$\mathcal{L}(x^*, y^*, \lambda) = (y+\lambda)(x+2\lambda) + \lambda(y+\lambda) + 2\lambda(x+2\lambda) - \lambda$
$= xy + 3\lambda y + 3\lambda x + 5\lambda^2 - \lambda$

$\nabla_\lambda \mathcal{L}(x^*, y^*, \lambda) = 0 = 3y + 3x + 10\lambda - 1$.
$\therefore \lambda = -\frac{1}{10}(3x + 3y - 1)$
$\therefore$ OPTIMAL x, y are
$x^* = y - \frac{1}{10}(3x + 3y - 1)$
$y^* = x - \frac{1}{5}(3x + 3y - 1)$

* Please refer sheet 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?

K-means clustering is a form of unsupervised learning. It aims to cluster unlabelled data points ~~[illegible]~~ to achieve classification. The algorithm randomly selects 'K' data points as classes, and other data points are clustered to the nearest selected datapoint. Then the mean of each cluster becomes the new center of cluster, and the classification repeats, until no data point changes class with next iteration.

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?

Labelled data is split into training and validation set. Model is trained on training set, then validated ~~[illegible]~~ (i.e. performance is measured) on validation set. The validation loss is a optimistic estimate of the true loss, because there is no guarantee that the labelled data captures all quality of unlabelled ~~[illegible]~~ 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.

model	Pro	Con
Decision Tree	The decisions are easily explained	easily overfits to training data as the tree grows
SVM	the soft margin allows for mislabelled data	~~lots of hyperparameters~~ hyperparameter tuning make it engineering heavy.
NN	Performs well in variety of tasks.	difficult to justify/explain the decisions. ~~[illegible]~~ i.e. hidden layers

Name: Kosuke James Nishio

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

Backpropagation is commonly used to tune/optimize the weights in neural nets. When training your model, you can minimize the loss function by comparing output to true label, then backpropagate through all weights (predicted) affecting the model and performing gradient descent. Efficiency can be increased by using stochastic gradient descent, to reduce the required computations.

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?

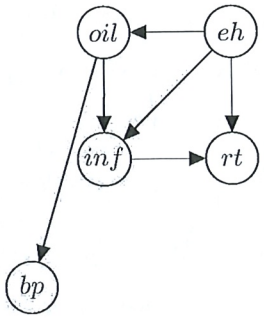
Convolutional layer gives the input data a useful structure, before it is fed into subsequent neural nets like FNN. In image classification, for example, a kernel is applied to the input image (i.e. convolution), detecting abstract structures like lines and patterns. The input sequence has to be carried through to update kernel (i.e. pixel location), which is different from standard back propagation.

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 deep learning can be analogous to human attention. For example, in Transformers performing text translation, it performs "self attention", in which every word in the input gets an attention score to every other word in the sentence. This is done through learning the key, value, query weights.

Name: Rosalie James Nyiro

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).



(c)

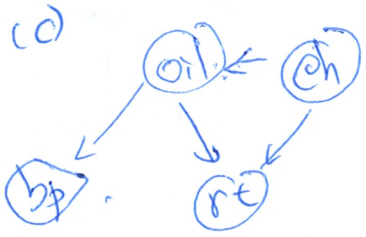


Figure 1: A simple Bayesian network

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. $P(oil, inf, eh, bp, rt) = P(oil|eh)P(inf|oil,eh)P(eh)P(bp|oil)P(rt|inf,eh)$
2. (a) No, 'bp' is dependent on 'oil' which is dependent on 'eh'.
   (b) Yes, 'oil' is dependent on 'eh', thus if 'oil' is observed $bp \perp eh$
   (c) No, 'oil' is unobserved creating a dependency.

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. $P(\theta|D) = \int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{F_\theta(t)-\mu}{\sigma}\right)^2} dt$.
2. Bayes can update the priors in light of new evidence.

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

$P(\theta)$ ← prior. $P(D)$ ← evidence

Universitat Pompeu Fabra Barcelona

Sheet 1

Nom i cognoms: Kosuke James Nishio.

Assignatura: K.Mro. Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Question 2

$$\max_{x,y}(xy)$$

s.t. $x+2y=1$

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

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

$$\therefore x^* = -2\lambda$$
$$y^* = -\lambda$$

$$\nabla_\lambda \mathcal{L}(x^*,y^*,\lambda) = 0 = \nabla_\lambda \left[2\lambda^2 + (-2\lambda)\lambda + 2\lambda(-\lambda) - \lambda \right]$$

$$= \nabla_\lambda \left[-2\lambda^2 - \lambda \right]$$

$$= -4\lambda - 1$$

$$\therefore \lambda^* = -1/4$$

$$\therefore x^* = 1/2, \quad y^* = 1/4$$

$$\max_{x,y} xy = 1/8$$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: DAVIDE LOCATELLI

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

We need to show that $f''(x) \geq 0$ to prove convexity

(1) For $f(x) = e^x$ we have that $f''(x) = e^x$ and since $x > 0$ we know that for any x e^x will be > 0
Hence $f''(x) > 0$ ■

(2) For $f(x) = x \log x$ we have that $f''(x) = \frac{1}{x}$
because $f'(x) = 1 \log x + x \cdot \frac{1}{x} = \log x$
Since $x > 0$ we know that $\frac{1}{x} > 0$ for any x
Hence $f''(x) > 0$ ■

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.

We have the equality constraint $x + 2y - 1 = 0$

[See extra sheet 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?

k-means aims to minimize the cost of assigning each point x to a cluster C_i of k clusters, where the cost is the squared distance between the point x_i and the centroid μ_i of the cluster. Mathematically, we want to minimize

$$\text{Total\_cost} = \sum_{i}^{k} \sum_{x \in C_i} (d(x, \mu_i))^2$$

To approximate the solution, the algorithm starts by picking k random centroids. Then, using the centroids, it computes the clusters by assigning each point to the closest centroid. Then it recomputes the centroids by minimizing its distance with all the points in the cluster. Repeat...

[see extra sheet ①]

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 a supervised learning problem we have a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ which is sampled from an unknown distribution D over our domain set X and an unknown labeling function $f: X \to Y$. A loss function $\ell: Y \times Y \to \mathbb{R}$ measures the error of an hypothesis function $h \in H$ s.t. $h: X \to Y$ and h approximates f. The loss function ℓ can be used on our training set $S, \hat{y}$ and in this case it's known as empirical risk, $L_S(\hat{y}, y)$. The true loss is the loss function applied to D, $L_D(\hat{y}, y)$. But since D is unknown, we cannot observe L_D. Hence, we can split our S into two smaller disjoint sets...

[see extra sheet ① and ②]

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:

⊕ we obtain an interpretable model: each node represents a decision split and given a leaf (i.e. a prediction label) we can go up the tree and observe which decisions the model based the prediction on

⊖ the shape of the tree is sensitive to the data observed in training, so we are likely to overfit. One can overcome this by random sampling of the data, generating a tree for each sampled subset, and combining the trees. Also can use pruning techniques.

[see extra sheet ②]

Name: DAVIDE LOCATELLI

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

Like in stochastic gradient descent, an objective is to minimize a loss function, i.e. reduce an model's errors. (*)
However, in neural networks all neurons in one layer will be connected to each of the neurons in the next layer. Hence, the gradient in the last layer needs to be propagated backwards to the first layer for us to be able to calculate the partial derivatives.
We can use the chain rule in order to efficiently calculate this gradient.

(*) We do this by calculating the gradient of the function and setting it to zero. The we move by $-\eta \nabla L(\hat{y}, y)$ where η is the learning rate.

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 applies a transformation to an input by means of a kernel. The kernel contains the transformation function, and the convolutional layer applies this kernel to the input by moving the kernel

[see extra sheet (2)]

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 is a mechanism that was introduced in computer vision which shows where a network "attends to" (i.e. focuses on) when classifying an object. Given that the model predicts that there is an object it attention shows which parts of the image influenced the decision, i.e. where the network thinks it is.

K = house attention

[see sheet (3)]

Name: Davide Scatelli

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).

rt ← eh → oil → bp
eh → inf ← oil → bp

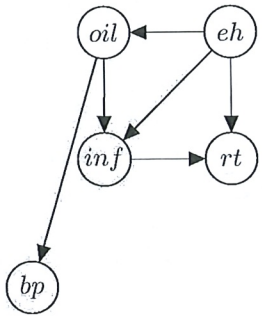


Figure 1: A simple Bayesian network

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?

p(eh, oil, inf, rt, bp) = p(eh) · p(rt|eh)p(eh) · p(inf|oil, eh)p(oil)p(eh) · p(bp|oil)p(oil) + p(oil|eh)p(eh)
Basically if a node has parents, they will be in the conditional probability and we multi
[see extra sheets]
(2) and (3)

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: Davide Locatelli

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data: 14 Dec 2021

(1) Workings of second derivative of $x \log x$

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

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

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

---

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

s.t. $x + 2y = 1$

We have the equality constraint $x + 2y - 1 = 0$

The lagrangian is

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

$= xy + \lambda x + 2\lambda y - \lambda$

We set the gradient to zero

$\delta_x \mathcal{L}(x, y^*, \lambda) = 0$

$y + \lambda = 0$

$\boxed{y = -\lambda}$

$\delta_y \mathcal{L}(x^*, y, \lambda) = 0$

$x + 2\lambda = 0$

$\boxed{x = -2\lambda}$

To find λ^* we have

$\lambda^* = \min_{\lambda} g(\lambda) = \mathcal{L}(x^*, y^*, \lambda)$

we set the gradient to zero

$$g(\lambda) = (-2\lambda)(-\lambda) + \lambda(-2\lambda) + 2\lambda(-\lambda) - \lambda$$

$$= 2\lambda^2 - 2\lambda^2 - 2\lambda^2 - \lambda$$

$$= -2\lambda^2 - \lambda$$

$$\frac{\partial}{\partial \lambda} g(\lambda) = -2 \cdot 2\lambda - 1 = -4\lambda - 1$$

Set to zero

$$-4\lambda - 1 = 0$$

$$-4\lambda = 1$$

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

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

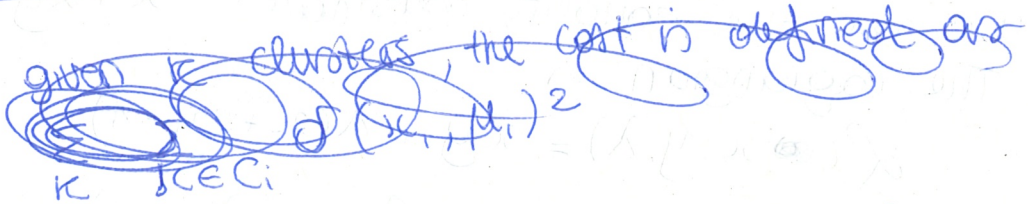
Now we can find x^* and y^*

$$x^* = -2\left(-\frac{1}{4}\right) = \frac{1}{2}$$

$$y^* = -\left(-\frac{1}{4}\right) = \frac{1}{4}$$

---

(3)



...until the centroids do not move anymore.
This is an instance of Expectation-Maximization.

---

(4) ... one of which we will call S_1, and we will use it for Empirical Risk minimization, and the other we will call V, and we will use this set as a validation set on which to test our $h^* = ERM_{H}(S_1)$.
We will calculate the validation loss $L_V(\hat{y}, y)$ by using our h^* to predict a $\hat{y}$ and comparing that with the y in our validation set.
L_V will be a good estimate of L_D because set V was sampled from S which was in turn sampled from D.

upf. Universitat Pompeu Fabra Barcelona

Nom i cognoms: Davide Locatelli

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data: 14 Dec 2021

(4 cont) Using L_T we can get an idea of the generalization ability of an model.

---

(5) Support Vector Machines

(+) Can be used on data that appears to be not linearly separable by means of kernel tricks

(−) learns hard boundaries (hyperplanes) that can misclassify the data at test time due to overfitting. Can use soft-SVM to overcome this.

Neural Networks

(+) can approximate any mathematical function including non linear ones

(−) each neuron is connected to the others in next layer meaning that there is higher computational expenses in the training process since we need to do backpropagation to minimize loss. Also, there are many hyperparameters that need to be chosen.

---

(7) A convolutional layer makes use of a kernel to apply a transformation to the input. If the input is an image, we can see the image as 3 matrices where cells indicate pixel values for red, blue and green. A kernel will be a smaller set of 3 matrices that contains values used to transform the image. The kernel is applied to a subsection of the image and then moved by a certain stride. The resulting transformed smaller matrices capture features of the

original input such as vertical, horizontal, diagonal lines or presence of certain higher level features depending on the transformation applied.

, the neurons in a convolutional layer share their weights, i.e. the values inside the kernel. Hence our backpropagation algorithm needs to be adjusted to account for weight sharing.

---

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

(1)

(2) a) $eh \perp\!\!\!\perp bp$?

We have path

$eh \rightarrow oil \rightarrow bp$

It's a head-to-tail path and oil is not part of our evidence, so the path is unblocked

Hence $eh \not\perp\!\!\!\perp bp$

b) $eh \perp\!\!\!\perp bp \mid oil = high$?

We have path

$eh \rightarrow oil \rightarrow bp$

It's head-to-tail and oil is part of our evidence, so the path is blocked

Hence $eh \perp\!\!\!\perp bp \mid oil$

c) $rt \perp\!\!\!\perp bp \mid eh$?

We have paths

I. $rt \leftarrow eh \rightarrow oil \rightarrow bp$ tail-to-tail and eh part of evidence so blocked

II. $rt \leftarrow eh \rightarrow inf \leftarrow oil \rightarrow bp$ head-to-head on inf and inf not part of evidence so blocked

III $rt \leftarrow inf \leftarrow oil \rightarrow bp$ UNBLOCKED! head-to-tail and tail-to-tail where inf and oil are not part of evidence

upf. Universitat Pompeu Fabra Barcelona

Nom i cognoms: DAVIDE COCATELLI

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

9 cont) since not all paths are blocked

---

8) ... this mechanism was integrated in NLP with the introduction of Transformers. In this context, attention is a weighted score for each word in the sentence given the current word.
As an illustrative example, in BERT (a transformer based architecture), we have the goal of language modeling. Take the sentence

"Jack sees Aaron's guitar, but he won't pick it up"

what does "he" refer to? Jack, Aaron or the guitar? Typically, it will refer to "Jack". So, BERT will make use of the dot product to obtain the word relatedness and use this calculation in its attention mechanism to attend to "Jack" when it's processing "he".

10)

In a forward probability problem we are after the probability of the data given that we have a generative model.

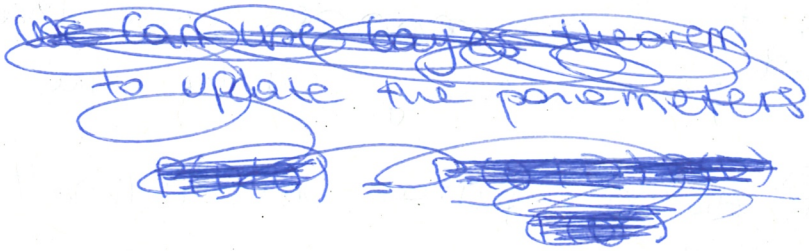
i.e. $P(D|\theta, H)$

In an inverse probability problem we are after the probability of the hidden variables of a model, given the observed data, where hidden variables can be the parameters of the model

i.e. $P(\theta|D, H)$.

Here we are after the likelihood of the parameters θ given fixed data D.

Here we are after $P(D|\theta)$



1) Since likelihood is not a function of the data but of the parameters $P(D|\theta)$ cannot be named likelihood of data given the parameters, but instead is called likelihood of the parameters given the data.

2) We can use Bayes' to update the parameters

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

we know that the density of a gaussian

upf. Universitat Pompeu Fabra Barcelona

Nom i cognoms: Davide Locatelli

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data: 16 Dec 2021

(10 cont.) distributed variable y is

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

So we know that $P(D|\theta) = f(y)$ since $F_\theta(t)$ will output σ^2 and μ depending on θ.

We can assume a uniform distribution for the prior, i.e. all $P(\theta)$ are equally likely

We will then pick the θ^* with the highest probability.