Machine Learning 2021-22
Final Exam

14 December 2021

Name: David. Arnau. Blasco

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:

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

  Knowing that $e^0 = 1$ and given that $x > 0$, we can say that the function is convex.

- $x \cdot \log x$

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

  Given that $x > 0$, $\frac{1}{x} > 0$. So we can assert that the function 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.

$\frac{\partial}{\partial x}$ $x+2y \rightarrow 1+2y=1 \rightarrow y=0$ | $x+2\cdot 0 = 1 \rightarrow x = 1$

$\frac{\partial}{\partial y}$ $x+2y \rightarrow x+2=1 \rightarrow x=-1$ | $-1+2y=1 \rightarrow y=1$

$\max\{0 \cdot 1, -1 \cdot 1\} = 0 \rightarrow$ optimal value

$x = 1$, $y = 0$ { optimal values of x, y

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

k-means clustering is used in unsupervised learning. Its objective is to find patterns in unlabeled data. Particularly, k-means algorithm is a centroid-based approach which means that clusters are defined depending on a center. This center is computed in each iteration using the mean distance between the k-nearest points. Finally, we are able to plot the data in (not always) well-defined groups which indicate that represent one pattern (or class).

where • are centroids

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 dataset helps to know if the model has been correctly trained or if it suffers overfitting or underfitting. This is a previous set to test set which tries to identify problems in the training.
- The fact of using a previous validation set gives us an idea how the model is performing. This is why we can say that the validation loss is an estimate of true loss. However, we cannot be sure the validation set to be different of training set, and in that case, we would not be able to detect if model suffers overfitting (we are using the same data). This is why we say it is optimistic.

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.

- Decission trees: easily trazable. Not good when we have a lot of data.
- SVM: very good to solve linear problems. When data is not separable we need to use kernel trick which increases the computational cost.
- Neural networks: Perfect to solve complex model. As it solves complex models its workflow is also complex, so this fact can produce some problems (overfitting, underfitting, gradient vanishing...) that must be controlled (usually with hyperparameters). In addition it requires a lot of training. and it's not trazable. depending on which neural network

Name: David Arcas Blasco

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

Backpropagation algorithm tries to update weights in a NN. What it does is to backpropagate the weights to the neurons of previous layers and makes corrections using gradient descendent.
I don't remember how it was exactly. I guess it was something like this: $W_i \leftarrow w_0 - \eta \, y_i W^T x_i$ Where η is the learning rate which is a crucial hyperparameter in order to find a minimum. Lastly.

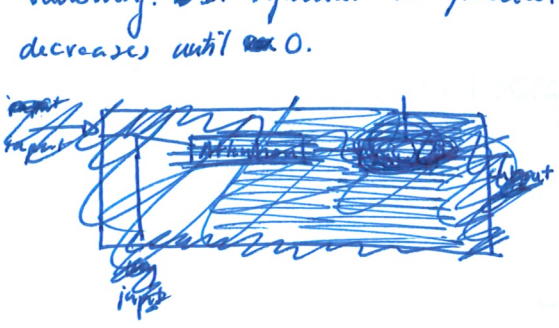
Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?

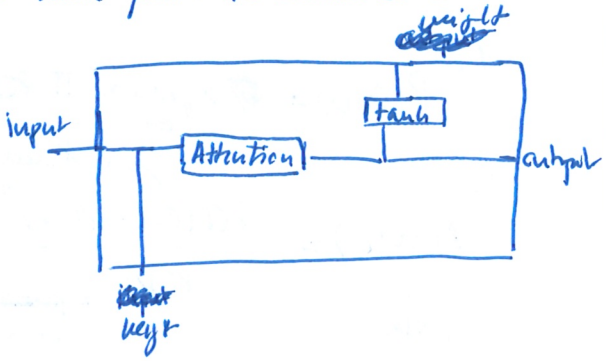
Convolutional layer tries to extract the most relevant features of a certain image. Basically, it is a very restricted filter (in a matrix form) that tries to obtain certain relevant features, in order to process the image.

In practice these filters are obtained using backpropagation algorithm which makes corrections in the filters. They are matrix with several 0 and some values that are those that will try to extract the feature.

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

The attention mechanism tries to avoid gradient exploding and gradient vanishing. It regulates the gradient when grows too much or when decreases until 0.





Name: David Arnau Blasco

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).

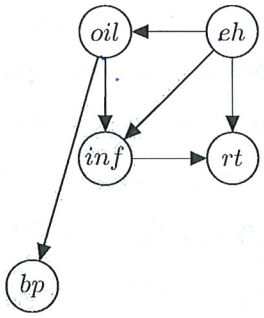


Figure 1: A simple Bayesian network

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. $P(x) = P(eh)\, P(oil|eh) \cdot p(rt|eh, inf) \cdot p(inf|oil, eh) \cdot p(bp|oil)$

2. a) $eh \perp\!\!\!\perp bp\,|\,\emptyset$? Paths: $eh \to oil \to bp$. Non blocked so $eh \not\perp\!\!\!\perp bp\,|\,\emptyset$

   b) $eh \perp\!\!\!\perp bp\,|\,oil$? Paths: $eh \to oil \to bp$. Blocked so $eh \perp\!\!\!\perp bp\,|\,oil$

   c) $rt \perp\!\!\!\perp bp\,|\,eh$? Paths: $rt \leftarrow eh \to oil \to bp$. Blocked
   $rt \leftarrow inf \leftarrow oil \to bp$ Non blocked so: $rt \not\perp\!\!\!\perp bp\,|\,eh$

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

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

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

1. $P(\theta|D) = \arg\max_\theta \prod P(D|\theta)$

2. likelihood, prior

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

posterior; evidence

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Jordan Harris

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

$\frac{d}{dx} e^x \cdot x\log x = e^x \cdot x\log x + e^x \cdot \frac{1}{x}$

non-negative

$x \neq 0$

$e^x > 1/x$

$\therefore$ 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.

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

$f(x) = y + \lambda$

$f(y) = x + 2\lambda$

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

$\lambda = 1/8$

$x = -1/4$

$y = 5/8$

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 object of K-means clustering is to assign input data into distinct groups as classifications. doing so by best approximating a centroid of ~~the~~ each cluster. (The most apple-ish of all the apples)

- The algorithm is first given a number of groups to look for (k)
- Then it seeks to maximize the distance between centroids & minimize the distance between those of the same group.
- Then it reevaluates which node is the best centroid, reassigns them logic.

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 mixes in the training set with the test set & creating one large group. It then partitions it arbitrarily. Then learns.

This allows for a greater generalization of the true hypothesis space. (Trained on the real world)

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: Very computationally safe & fast to converge. but may have an issue w/ under fitting.
- SVM: Great for linearly solvable problems, almost always converges, but issues with speed (slow) & complexity
- NN: Great at handling highly complex problems w/ lots of features. But there is a huge difficulty when trying to solve a problem w/ higher dimensions.

Name: Jordan Harris

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

Backpropegation can be implemented on a FFNN. that is a cyclic. (input → output)

In summary, while learning the algorithm is able to "trace" back down the network to "see" what sequence of gates led to a particular decision. This acts as a sort of memory! When a mistake is encountered it can then go back & find a better route.

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 NN's are fully connected between each layer. This allow's for each layer to hyper focus & rearrange at each layer to solve different aspects of the problem. ("each layer has diff types of approches to the problem")

- This ambiguity of connection due to the cyclic nature, in addition to the specificity of each layer makes backpropegation ineffective. (i.e. you would unlearn what the layer knew.

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 the ability of a network to monitor the traffic through & connectedness of each node. This informs what parts of the network are most important & what nodes may need to be dropped.

Name: Jordan Harris

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).

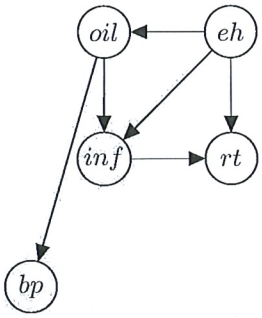


Figure 1: A simple Bayesian network

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

2a.) Yes, because the path (bp > oil > inf > rt) is blocked

b.) No, because the path from (bp > oil > eh) is open

c.) Yes, because of the fish tail, oil ← eh → 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 $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$.

P

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

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Ilse Meijer

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

A function f is convex when $f''(x) \geq 0$.

$f(x) = e^x$

$f'(x) = e^x$

$f''(x) = e^x \geq 0 \quad \forall x \in \mathbb{R}$

$f(x) = x \log x$

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

$f''(x) = \frac{1}{x} \geq 0 \quad \forall x > 0$

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.

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

$\nabla_{x,y} \mathcal{L}(x^*, y^*, \beta) = \begin{pmatrix} y^* - \beta \\ x^* - 2\beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$y^* - \beta = 0 \Rightarrow y^* = \beta$

$x^* - 2\beta = 0 \Rightarrow x^* = 2\beta$

$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \beta \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

We can derive and solve the dual problem

$\max_\beta \mathcal{L}(x^*, y^*, \beta) = 2\beta \cdot \beta + \beta(1 - 2\beta - 2 \cdot \beta) = -2\beta^2 + \beta$

$\nabla_\beta \mathcal{L}(x^*, y^*, \beta) = -4\beta + 1 = 0$

$\beta = \frac{1}{4} \Rightarrow \begin{pmatrix} x^* \\ y^* \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

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

The objective of k-means clustering is to find k clusters of nodes that are close to each other (by some distance metric)

The algorithm initializes with a random clustering of k-clusters and in each step it first recalculates which elements belong to which cluster given the means of the clusters and taking the cluster with minimum distance (see extra)

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 validation you separate your data in k-clusters and train your model on k-1 of these clusters and validate it on the k^{th} cluster. You repeat this until every cluster is used exactly once as the validation cluster. The validation loss is the average loss over all these validations. It is optimistic since you use each cluster for both training and validation (at least once)

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: ⊕ easy to interpret
⊖ requires a lot of engineering (too big a tree can overfit, too small won't tell anything)

SVM: ⊕ can by using the kernal trick it can also classify non-linearly separable data without extra computational power.
⊖ Difficult to interpret (especially when using the kernal trick)

NN: ⊕ feature design is handled by the model
⊖ less Difficult to interpret
(⊖ computationally heavy)

Name: Ilse Meijer

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

Backpropagation is used to update weights. When a mistake is found the weights are updated (e.g. in linear regression by: $w_1 = w_0 + y_i x_i$ (x_i, y_i mistake) this is repeated until convergence.

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

In a convolutional layer the ~~nodes~~ layers is ~~are~~ not fully connected and the weig some of the weights are shared. It is often used when processing images, where a ▦-filter is used to process the image (in a more abstract way for each additional convolutional layer). For backpropagation we can change the weights in the filter, but not specifically for just one output ~~(it changes in every one)~~ since the filter is shared.

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 can focus on some of the elements in the context of a datapoint more than on others by some probabilistic relation in the training data. In a transformer network multiple self-attention layers are incorporated, together with hidden layers as in neural networks ~~layers~~ and e.g. a softmax.

Name: Ilse Meijer

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).

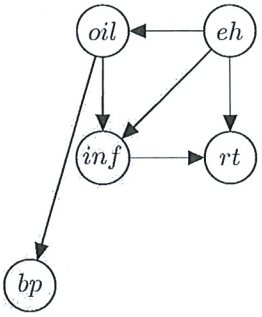


Figure 1: A simple Bayesian network

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

2a) $eh \not\perp bp \mid \emptyset$? No, because the path trough oil is not blocked, since oil is a head-to-tail node and is not in the evidence. Thus dependent.

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

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

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

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

2. Bayes theorem: $P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}$

$= P(D|\theta) = \frac{P(\theta|D)P(D)}{P(\theta)}$ we know the distribution for D, so $P(D) = 1$

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Ilse Meijer

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

3) between the mean of that cluster and the datapoint. for each datapoint.
It then Recalculates the mean of each cluster given the ~~mean~~ adjusted clusters.
It repeats this till convergence.

g.2.b) eh ⊥⊥ bp | oil=high?
Yes, since every path has to go through the node oil, which in every path is either a head-to-tail or tail-to-tail node. Because oil is in the evidence, we thus know that every path is blocked. Thus independent

c) At ⊥⊥ bp | eh=low?
No, since the path through oil and int is not blocked: oil is a tail-to-tail node and int is a head-to-tail node. Neither of them are in the evidence. Thus dependent.

10) $P(D|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta - F_\theta(t)}{\sigma}\right)^2}$

Assume $P(\theta) = 1$

$$P(D) = \int_0^1 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta - F_\theta(t)}{\sigma}\right)^2} d\theta$$

Thus $$P(\theta|D) = \frac{e^{-\frac{1}{2}\left(\frac{\theta - F_\theta(t)}{\sigma}\right)^2}}{\int_0^1 e^{-\frac{1}{2}\left(\frac{\theta - F_\theta(t)}{\sigma}\right)^2} d\theta}$$

$$P(y|D) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y - F_\theta(t)}{\sigma}\right)^2}$$

$$= \int_0^1 P(D|\theta) P(\theta) \, d\theta$$

Machine Learning 2021-22
Final Exam

14 December 2021

Name: Charles 'Dean' Cochran

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

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

$f(x) = e^x$

$f'(x) = e^x \Rightarrow$ continuous

$f''(x) = e^x \Rightarrow > 0$

Since $f'(x)$ is continuous and $f''(x)$ is non-negative, $f(x)$ is convex

$a(x) = x \quad b(x) = \log x$

$g(x) = a(x) \cdot b(x) = x \log x$

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

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

Since $g'(x)$ is continuous, $g''(x)$ is non negative, $g(x)$ is convex

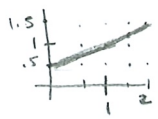
Question 2: [1 point] Solve the following constrained optimization problem:

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

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

$x + 2y = 1$

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



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

$L_x = 1 + \lambda = 0 \Rightarrow \lambda_x = -1$

$L_y = 2 + 2\lambda = 0 \Rightarrow \lambda_y = 1$

$x^* = -1$

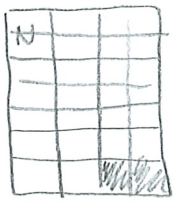
$y^* = 1$

$xy = (1)(-1) = -1$

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 objective of k-means clustering is to group instances with similar feature values. K-means approximates this by classifying instances to the closest cluster, It measures the proximity to the other clusters w/ various distance metrics, most commonly Euclidean Distance. This algorithm is calculated over and over again, adjusting the centroid every iteration until all instances are correctly classified and the instances need no adjustment

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?



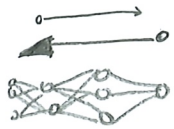
$L_{CV}(\phi) = \frac{1}{k}\sum_{i=1} L_{S_i}(h_i)$ Cross validation is one instance of verifying a model inorder to more accurately measure an estimate of true loss. This is an optimistic estimate because it provides an accumulation of all loss functions for each model instance. This returns a more valuable estimate than a singular loss estimation.

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.

Supervised Learning	Advantage	Disadvantage
Decision Trees	Easily interpretable,	isn't granted to find global min, can overfit model in high complexities
Support Vector Machines	very flexible and can be utilized in different feature spaces thanks to the kernel trick	highly dependent on parameter selection C, kernel, regularization, etc..
Neural Networks	Can approximate any continuous function given the correct structure	A lot of interpretability is lost, and it is also highly dependent on hyperparameter selection η, T, size of batches, size of layer etc..

Name: Charles 'Dean' Calvan

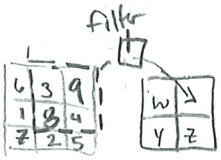
Question 6: [1 point] Briefly explain the backpropagation algorithm, and describe how the algorithm can be efficiently implemented.



$-\eta d_k X_{t-1}$

Backpropagation is an algorithm that is used in neural networks to adjust the weights after a prediction has been made. The Algorithm uses a learning rate to adjust how much adjustment is needed to be applied for every calculation. A d_k functin Calculates the necessary gradients and matrix operations given the previous input of X_{t-1}. All of this is done for each node in every layer. This algorithm is often adjusted when discussing RNN's to limit the backpropagation.

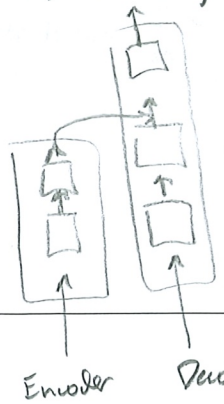
Question 7: [1 point] Explain how a convolutional layer works in deep learning. How are convolutional layers implemented in practice? What changes do we have to make to standard backpropagation?



Convolutional Neural Networks use convolutional layers in their 'non-fully-connected' structure to extract low-level feature values. These are best expressed as different types of filters which the model can use to dimensionally reduce the scale of the neural networks data. This in turn allows for the model to generalize large portions of our data and still manage to retain the data it learned. In practice the are many convolution layers implemented for image processing meaning filters can be applied in array matricies of 2 or 3D shape. When backpropogation Matrix Operations also must interpret filtered data to produce estimates of the original data. This is process is maximized by minimizing information loss. Additionally height matricies are shared.... continued on seperate paper...

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

Attention is a mechanism that accepts information in many different scenarios. Wether inside an encoder or decoder the attention plays a large role in RNN, and transformer networks.



Attention is used in both ecoding and decoding in order to assist the model when it needs to prioritize its learning capabilities.

Name: Charles 'Dean' Cochran

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).

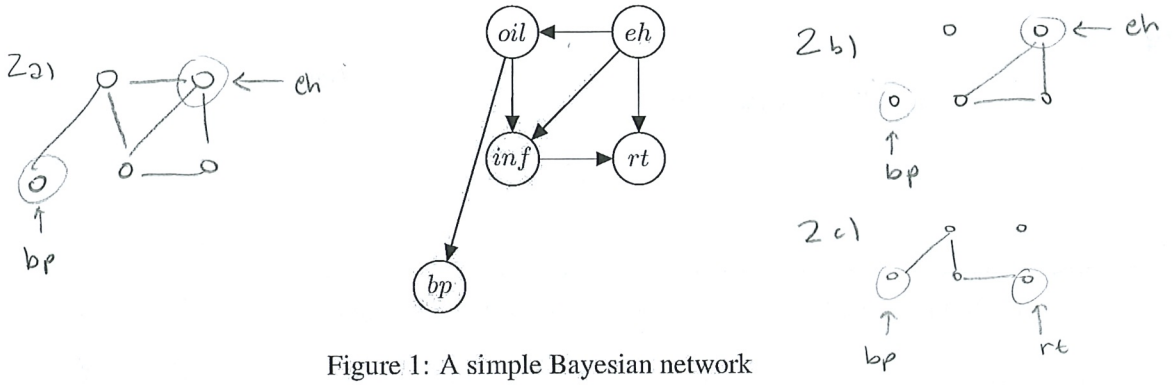


Figure 1: A simple Bayesian network

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) $P(oil, inf, eh, bp, rt) \propto \phi f_1(eh, oil)\, \phi f_2(oil, bp)\, \phi f_3(oil, inf)\, \phi f_4(eh, inf)\, \phi f_5(eh, rt)\, \phi f_6(inf, rt)$

2a) $eh \perp\!\!\!\perp bp \mid \emptyset$
False, there exists a path through 'oil'
$eh \rightarrow \{rt, inf, oil\} \rightarrow bp$

2b) $eh \perp\!\!\!\perp bp \mid oil = high$
True, there doesn't exist a path from eh to bp
$\{\emptyset\}$

2c) $rt \perp\!\!\!\perp bp \mid eh = low$
False, there exists a path from rt to bp
$rt \rightarrow inf \rightarrow oil \rightarrow bp$

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

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

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

Bayes I: $P(\theta \mid D, H) = \frac{P(D \mid \theta, H) P(\theta \mid H)}{P(D \mid H)}$

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

$y \sim N(\mu, \sigma^2)$

Bayes II: $P(H \mid D) = \frac{P(D \mid H) P(H)}{P(D)}$

Bayes theorem illustrates that by maximizing likelihood you are also minimizing loss for Gaussian generative models

Assuming θ exists, and we are given $\mathcal{D}$ we can use Bayes I and Bayes II to calculate the necessary posteriors by finding the $\frac{\text{likelihood} \times \text{priors}}{\text{evidence}}$

In this case our posterior is $p(\theta \mid D, H)$, likelihood is $p(D \mid \theta, H)$, prior is $p(\theta \mid H)$ evidence is $p(D \mid H)$

Universitat Pompeu Fabra Barcelona

Nom i cognoms: Charles 'Dean' Calvan

Assignatura: Grup: Curs:

Centre/Estudi:

Professor/a:

Data:

Question #7 Continued The Sharing of matrices forces us to alter the backpropagation accordingly (Though this diminishes the computation required)