Review on May 1, 2023

for May 5, 2023, 2p Final Exam

last updated; May 1, 10a

1) Belief Networks and Naïve Bayes

a) Consider the following belief network that consists of variables A, B, C, D, E all of which have two states {true, false} and whose structure is depicted below is given.

 B

A D E

 C

1. Is A|∅ d-separable from E|∅[[1]](#footnote-1). Given reasons for your answer! [4]

There are two paths:

A-C-D-E and A-B-D-E

The first path is blocked in node C (pattern3 as C in not in evidence) and

The second path is blocked in B (pattern3 as B is not in evidence)

As both paths are blocked, A and E are independent

1. Is B|A d-separable from C|A. Give reasons for your answer! [4]

There are two paths: B-D-C and B-A-C

The first path is not blocked in node D, as D is not in evidence

The second path is blocked in A (pattern2 as A is not in evidence) if they do not write that line they still deserve full credit

As the first path is not blocked B|A is not d-separable (independent) from C|A

1. What advantage you see in using Belief Networks instead of using a naïve Bayesian approach? [4]

Using Belief-networks you can express dependencies between random variables that are not independent of each other, using domain specific knowledge. By doing that BBNs will obtain “better”, more accurate predictions than the naïve Bayesian approach, as making conditional independence assumptions that are violated by the observed data will increase prediction errors.

1. Assume P(D)=0.02, P(S1)=0.2 P(S1|D)=0.4 P(S2)=0.1 P(S2|D)=0.3. Compute P(D|S1,S2) using a Naïve Bayesian approach! [3]

P(D|S1,S2)=0.02\*2\*3=0.12 No partial credit!

~~2) AI in General [7]~~ not relevant

~~Recently, foreign governments have assigned a lot of resources to AI education, research and development. For example, the Chinese Government wants “~~*~~AI to be Made in China~~*~~” by 2030. What do you believe are the reasons for this development? Limit your discussion to 5-8 sentences!~~

~~No Answer given! Expect a different essay-style question in the final exam as the review material is different!~~

3) All kind of Questions

a) Assume P(A|B) is 0.2 and P(B) is 0.7 and P(C|A,B) is 0.1; compute P(A,B,C) (this is the same as P(A∧B∧C))

P(A,B,C)=P(B)\*P(A,C|B)=P(B)\*P(A|B)\*P(C|A,B)=0.7\*0.2\*0.1=…

Using P(A,B)=P(A)\*P(B|A)=P(B)\*P(A|B)

b) SVMs are often using in conjunction with Kernel functions φ to learn to classify 2 classes based on a training set D; how is the hyperplane obtained in this case?

1) D’=φ(D)

2) Learn a hyperplane for φ(D)─in the “mapped” space

c) Assume a support vector machine hyperplane that has been learnt for a dataset having attributes A, B, C is given:

A2+B3-C4-2

e.g. the support vector machine classifies an example ex1 (0,1,1) where 0, 1, 1 are the values for attributes A, B, C as belonging to the negative class as 0\*2+1\*3-4\*1-2=3.

Assume we have 3 more examples ex2, ex3, ex4 are given for which the hyperplane equation returns -20, 0, +4. What does this tell you about the examples [3].

ex1and ex2 are on the side of the negative class of the hyperplane[1], but ex2 is much further away from the hyperplane than ex1[0.5]; ex3 is located on the hyperplane[1], and ex4 is one the other (positive class) side if the hyperplane [0.5]

d) Can the fact that the support vector machine equation returns a negative or positive number—and not just a class label of the predicted class—be used for anything useful?

If the values returned for the hyperplane equation for a testing example e is very close to 0 we are much less confident about the correctness of the SVM’s prediction compared to the case where plugging the attribute values of e into the hyperplane equation returns large negative or positive values.

e) Assume you train a model (e.g. a neural network or decision tree model) and you observe overfitting; what can be done to alleviate the problem?

1. reduce model complexity (e.g. use less nodes / less layers in the case of a neural network
2. increase size of training sets

f) What role do validation sets play in Supervised Learning? [3]

It is used during training to determine the optimal setting for the employed machine learning algorithm’s (hyper) parameters.

If the do not mention ‘during training’ only 2 points!

g) Neural networks that use at least one (or a lot) intermediate layers have been much more successful than 2-layer neural networks which directly link inputs with outputs. Why do you believe this is the case? [3]

The intermediate layer allows to create new features that facilitate getting high accuracies [3].

Other answers might deserve partial credit to up to 2 points.

h) Why did generators—e.g. obtained using a variational autoencoder—gain a lot of importance recently:

a. Can create novel design, novel faces, novel stories,…

b. Can be used to augment train sets to alleviate the problems of overfitting

4) Hidden Markov Models

1. What HMM are mathematically: λ = (A, B, Π), parameter set of HMM

2. Take a look at the weather activity HMM in <https://en.wikipedia.org/wiki/Hidden_Markov_model>

3. What kind of questions can be answered using HMMs?

C



Assume the above hidden Markov Model C is given (taken from the HMM Wikepedia page) with states Rainy (R for short) and Sunny (S for short) and outputs Walk (W for short), Shop (S for short) and clean (C for short).

1. What does the edge labled with 0.4 going from the node Sunny to node Rainy mean? [1.5]

The probability of going from state Sunny to state Rainy in the next step is 0.4.

1. What does the edge labeled with 0.6 going from node Sunny to node Walk mean? [1.5]

The probability of outputing Walk when in state Sunny is 0.6.

1. What kind of problems can be solved with HMMs? Also mention one important application of HMMs. [4]

HMMs allow the prediction of state sequences/being in a single state of a probabilistic state space (Markov chain) based on observed output sequences whose probabilities to occur are state dependent. [3]

HMM allow to infer the likelihood of hidden state sequences (in probabilistic state space) based on observed outputs sequences whose probabilities to occur are state dependent.

 In applying HMM, a sequence is modelled as an output of a discrete stochastic process, which progresses through a series of states that are ‘hidden’ from the observer.[2]

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process — call it {\displaystyle X}X — with unobservable ("hidden") states. As part of the definition, HMM requires that there be an observable process {\displaystyle Y}Y whose outcomes are "influenced" by the outcomes of {\displaystyle X}X in a known way. [2]

Applications: Gene Sequence Analysis, speech recognition, robot localization [1]

**5) Neural Networks**

a) Describe how multi-layer neural networks, consisting of 3+ layers learns a model for a training set! Limit you answer to at most 9 sentences! [7]

Neural network learning tries to find weights that minimize the error in the neural network prediction for a training set [1]. Neural networks employ gradient decent hill climbing to find the “best” weights. [1]. In particular, Neural network learning adjust weights using the gradient of the error function of the training set [1]; the search starts with a random initial weight vector and weights are adjusted in the direction of the steepest negative gradient of this error function---that is weights are updated accordingly moving in the direction that reduces the error the most [2]; The step width of this weight update depends on the learning rate and other factors [1]. In order to apply this procedure the error for each none-input node has to be known. As this error is not initially given intermediate for intermediate layer nodes, it is computed using the back-propagation algorithm [2].

Other observation might deserve credit. At most 7 points!

b) Looking at the sub neural network in the figure below; what does the back-propagated associated error ΔA for a node A depend on? Give a formula to compute its value, assuming g is the activation function of the neural network.

 wA,B=0.5

 ΔB=0.4

 A B

wC**,A**=0.6 wD,A=0.2

 C D

ΔA depends on: the associated error ΔB in the node B, the weight of the connection between A and B and the derivative g’ of the activation function g for the linear input of node A zA.

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.5\*0.4=g’(0.12+0.08)\*0.5\*0.4=g’(0.2)\*0.5\*0.4

We assume that g(0.2)=0.2

**6. AutoEncoders**



A standard Autoencoder

What is the loss function for auto encoders? How is it different from the loss functions used by neural network for prediction and classification tasks?

The entire network of an autoencorder is usually trained as a whole. The loss function is usually either the mean-squared error or cross-entropy between the output and the input, known as the reconstruction loss, which penalizes the network for creating outputs different from the input.

Loss functions of prediction and classification tasks on the other hand compute the differences between the computed output and the expected output for training examples.



7) Online Credit Problem **i** Group P revisited

For the Burgulary-Earthquake-Alarm-John\_Calls-Mary\_Calls Belief Network in our textbook compute:

1. P(Alarm=Yes|Earthquake=Yes)
2. P(Mary\_Calls=Yes|Burgulary=No) too difficult for final exam!

Annotate every step of your computations (e.g. “Bayes Theorem”, “Definition of P(A,B)”), mention not obvious assumptions your computations made (e.g. “as A|F is d-separable for B|F” P(A,B|F)=P(A|F)\*P(B|F)…”).

P(Alarm|Earthquake)= P(Alarm,Earthquake)/P(Earthquake)=**?/**0.02≈0.3

**?=**P(Alarm,Earthquake)=P(Alarm,Earthquake,Burglary) **+** P(Alarm,Earthquake,~Burglary)

P(A, E, B)= P(A)\*P(B,E)\*P(A|B,E)= as B and E are d-separable given no evidence ∅

P(B)\*P(E)\*P(A|B,E)=0.01\*0.02\*0.95

P(A, E, ~B)= P(~B,E)\*P(A|~B,E)= as B and E are d-separable given no evidence ∅

P(~B)\*P(E)\*P(A|~B,E)=0.99\*0.02\*0.29

1. P(Mary\_Calls|~Burglary)=P(M | ~B)

8. Short Questions about the EU Ethical and Responsible AI document

9. Q-Learning centering on Q-Learning, SARSA, exploration/exploitation, and very general RL-concepts such as policies, objectives of reinforcement learning, deep reinforcement learning (only basics)

Consider the following World called DEF is given:



1. Assume we apply Q-learning, assuming the agent starts in state 3 and applies the operator sequence **s-w-n-e;** assume theinitial q-values are all 0; what are the new q-values that will be computed by Q-Learning? Assume α=0.5 and γ=1 in your computations. [7]

No solution given; but there might be a problem in the exam (see GHC presentations which solved similar problems for PD Worlds)

b) What is the main difference between Q-Learning and SARSA? [3]

Q-learning updates the q-values assuming a greedy policy, whereas SARSA updates the q-values assuming the actual policy.

1. ∅ represents “no evidence”; question i basically asks if A and E are independent; that is, if P(A∧E)=P(A)\*P(E) [↑](#footnote-ref-1)