Review on May 3, 2021

for May 12, 2021 Final Exam

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!

v) Naïve Bayesian systems make the conditional independence assumption when for example computing P(D|S1,S2,S4). What assumptions are exactly made? What advantages do you see in the approach? What are the drawbacks of making the conditional independence assumption?

P(D|S1,S2,S3)= P(D)\*P(S1,S2,S3|D)/P(S1,S2,S3)≈P(D)\*P(S1|D)/P(S1)\*P(S2|D)/P(S2)\*P(S3|D)/P(S3)

We have to assume that S1|D, S2|D, S3|D are independent and if the exact probability of P(D|S1,S2,S3) needs to be computed we additionally have to assume that S1, S2, and S3 are independent. If Naïve Bayes is just used for a classification problem, it is not necessary to know P(S1,S2,S3) as we are only interested in know if P(D|S1,S2,S3) is larger/smaller than P(D’|S1,S2,S3); consequently the exact value for P(S1,S2,S3) is not relevant!

Advantage: low knowledge acquisition cost; it is not necessary to acquire a lot of probabilities which might be expensive.

Disadvantage: obtained probabilities might not be correct;

One problem of using BBN in many cases you have to “make up” probabilities you don’t know.

2) AI in General [7]

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!

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) Why is leave-one-out cross validation[[2]](#footnote-2) more popular in data science contests than 10-fold cross validation?

Leave-one out cross-validation has less bias, as exactly the same training set/test set pairs have to be used, making it more difficult to “cheat”.

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.

4) Naïve Bayes Classifier

Assume you have a classification problem involving 2 classes C1 and C2. Given a training example (‘tall’, ‘obese’, ‘high’) giving attribute values for attributes height, body-weight, and blood pressure: how does a naïve Bayesian classifier determine the class label of the example (‘tall’, ‘obese’, ‘high’)? [6]

1. Compute P(C1|height=tall,body-weight=obese,blood-pressure=high) and

P(C2|height=tall,body-weight=obese,blood-pressure=high)

2. …(give details how a “proxy” for 2 probabilities is computed)

3. Choose the proxy with the higher probability

For details how the proxy is computed see: <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

5) Hidden Markov Models

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

2. Forward-Backward Algorithms; particularly being able to compute αr(i) and βr(i)

**Go through the example in the HMM.pptx lecture!**

3. What do the Viterbi and Baum-Welch algorithm actually compute (will not ask for technical details how they compute it)

4. Application of HMM (Koller video)

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

**6) Neural Networks**

Describe how multi-layer neural networks, consisting of 3+ layers learn 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!



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)

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(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)
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)
2. In leave-one-out cross validation each example forms a fold; e.g. if the dataset contains 10,000 there will 10,000 one example folds. [↑](#footnote-ref-2)