Introduction to Artificial Intelligence

UNIZG FER, AY 2021/2022

Exercises, v1

10 Machine learning

1 (T) We're building a machine learning model for predicting the cinema audience numbers. We're considering three features: day of the week, movie genre, and the movie production budget. **Which machine learning algorithm would be suitable to use for this problem, and why?**

A A decision tree, because we have three features of equal information gain

B A neural network, because we're predicting a numeric value

C A naïve Bayes classifier, because we're predicting discrete values (whole numbers)

D A neural network, because the movie budget is a numeric feature

2 (T) The naïve Bayes classifier is called "naive" because the model assumes the conditional independence of features x_j within class y. Under this assumption, the class likelihood $P(x_1, \ldots, x_n|y)$ can be replaced by the product $\prod_{j=1}^{n} P(x_j|y)$. **What is the motivation behind introducing the conditional independence assumption?**

A The ability to generalize to unseen instances

B Achieving a higher accuracy of the model on the train set

C The ability to use non-binary features

D Preventing an underflow when multiplying probabilities

3 (C) Little Johnny has spent every summer of the last seven years learning a new programming language. He distilled his valuable experiences into a list *"A programming language I like"*, where he described each language with four features and recorded whether he liked that language ($y = 1$) or not ($y = 0$). This is how his list looks like:

i	Evaluation	Execution	Paradigm	Type checking	y
1	lazy	compiled	imperative	static	0
2	strict	interpreted	declarative	dynamic	0
3	lazy	compiled	imperative	dynamic	0
4	lazy	interpreted	hybrid	static	0
5	strict	interpreted	imperative	static	1
6	lazy	compiled	hybrid	dynamic	1
7	strict	compiled	hybrid	dynamic	1

This summer Little Johnny wants to eat and sleep a lot, and again learn a new language. He shortlisted a language $\mathbf{x}$ with the following characteristics: $\mathbf{x} = (\text{strict}, \text{interpreted}, \text{hybrid}, \text{dynamic})$. However, this time around Little Johnny would like to know up front whether he'd like the language, to avoid wasting the entire summer. Help Little Johnny by applying the naïve Bayes classifier with Laplace add-one smoothing to the above dataset. **What is the probability that Little Johnny will like programming language x?**

A 0.694 B 0.431 C 0.856 D 0.799

4 (P) We're building a Naïve Bayes classifier for classifying Twitter messages according to sentiment. The goal is to classify every tweet into one of three classes: positive, negative, or neutral. Each tweet consists of at most 280 words and we represent it as a binary feature vector. The vector has 5000 features, and every feature corresponds to one out of 5000 words from our vocabulary. If the feature is set to 1, then this means that the corresponding word occurred in the tweet, otherwise it is set to 0. E.g., if $x_{42} = 1$, then this means that the 42nd word from our vocabulary has occurred in the tweet. The training of this model boils down to estimating prior class probabilities and class likelihoods from a labeled sample. **How many probabilities need to be estimated for this model?**

A 1683 B 30003 C 569 D 31683

5 (T) The decision tree and the Bayes classifier are both supervised machine learning algorithms. However, these algorithms are quite different. **What is the advantage and what is the disadvantage of the decision tree over the Bayes classifier?**

A The advantage is that decision trees are resistant to small changes in the input dataset, but the disadvantage is that we assume conditional independence of features

B The advantage is that every example can be classified into more than a single class, but the disadvantage is that the decision tree can easily overfit

C The advantage is that we can prune the decision tree to prevent overfitting, but the disadvantage is that underflows can occur when calculating the probability

D The advantage is that we can better explain why an example is classified as it is, but the disadvantage is that we do not have the probability of a classification decision

6 (C) We have available a dataset for *"An unforgettable 2025 summer on the Adria, right after coronavirus pandemic"*. The dataset contains the following examples, each with four features and the target label y:

i	Destination	Island	Accommodation	Transport	y
1	Istria	no	private	car	yes
2	Istria	no	private	airplane	yes
3	Dalmatia	yes	hotel	car	yes
4	Dalmatia	yes	hotel	bus	yes
5	Kvarner Gulf	no	camp	bus	no
6	Dalmatia	yes	private	airpline	no
7	Istria	no	camp	car	no

Run the ID3 algorithm on this dataset. If, at any step, the information gain of two or more features is tied, choose the feature listed first in the table (the left-most feature). **How does this decision tree look like**?

A The root node is Accommodation, and its child node is Destination with information gain 0.918

B The root node is Accommodation, and its child node is Destination with information gain 0.251

C The root node is Destination, and its child node is Accommodation with information gain 0.918

D The root node is Destination, and its child node is Accommodation with information gain 0.251

7 (P) We're training a decision tree model on a dataset which, unfortunatelly but inevitably, also contains some noise. We are aware that the presence of noise can overfit the model, so we use cross-validation to prune the tree. We've split the data into a training set D_t and a validation set D_v, so that $D_t \cap D_v = \varnothing$. We then test a number of trees of varying depths, from $d = 1$ to $d = 42$. By observing the prediction errors of these trees, we conclude that the optimal tree depth is $d = 17$.

What exactly does this mean?

A That the error on D_v for the tree with $d = 17$ is lower than the error on D_v for trees with $d < 17$ and $d > 17$, but the error is likely larger on D_v than on D_t

B That the error on D_v for the tree with $d = 17$ is larger than the error for that tree on D_t, and also larger than the error for any other depth d when measured on D_v

C That the tree with $d = 17$ achieves the lowest error on D_t, while on D_v the tree always yields a larger error, with the maximum error for $d = 1$

D That the errors for trees with $d = 16$ and $d = 18$ are larger on both D_t and D_v, with error on D_v likely being larger than the error on D_t