CS 6780: Advanced Machine Learning **(Due: 02/12/19)**

Submission Assignment #1

Instructor: Thorsten Joachims *Name:* Molly Ingram, Julien Neves*, Netid:* msi34, jmn252

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- Please typeset your submissions in LaTeX. Use the template provided in `a1_template.tex` for your answers. Please include your name and NetIDs with submission. date and time on the first page.
- Assignments are due at the beginning of class at 2:55 pm on the due date on CMS.
- Late assignments can be submitted on CMS up to Sunday, Feb 17. This is also when the solutions will be released.
- You can do this assignment in groups of 2. Please submit no more than one submission per group.
- All sources of material must be cited. The University Academic Code of Conduct will be strictly enforced.

Problem 1: Version Spaces

(5+5+5+10=25 points)

Weather	*Food*	*Meeting*	*Attend?*
cold	full	idc	yes
cold	empty	no	no
hot	full	yes	yes
cold	full	no	yes
hot	empty	no	no

Table 1: Dataset $\mathcal{D}$

(a)

Weather	*Food*	*Meeting*	*Attend?*	h_A	h_B	h_C	h_D
cold	full	idc	yes	y	y	y	n
cold	empty	no	no	y	n	n	n
hot	full	yes	yes	y	y	y	n
cold	full	no	yes	y	y	y	y
hot	empty	no	no	y	n	n	n

Table 2: Dataset $\mathcal{D}$ with Predictions

(b) As we can see from Table 2, only h_B and h_C are in VS($\mathcal{D}$) because they are the only consistent hypotheses.

(c) With the new data point, only h_B is consistent so only h_B is in VS($\mathcal{D}$).

(d)

I have modified the notation of the problem for clarity. I am specifying that each $x_i \in \{0,1\}^{k+1}$, so there are k attributes and one output to predict. If I had 2^k data points and only $k-1$ attributes, I could not create a consistent decision tree. For example, the dataset $(X, Y) = \{(1,0),(0,1),(1,1),(0,0)\}$ does not have a consistent decision tree.

Let $k \in \mathbb{Z}_+$. Consider a dataset $\mathcal{D} = \{x_1, x_2, ..., x_{2^k}\}$ such that $x_i \in \{0, 1\}^{k+1}$ and $x_i \neq x_j \forall i \neq j$. Let X be the first k attributes of each datapoint and Y, the remaining attribute, be the output variable to predict. Consider a decision tree where the first decision node is (Attribute 1 == value), the second decision node is (Attribute 2 == value), and so on where the k^{th} decision node is (Attribute k == value). Since each attribute is binary, the tree ends with 2^k leaf nodes. Because $x_i \neq x_j \forall i \neq j$, each path down the tree corresponds to one datapoint, thus the decision tree is consistent and an element of VS($\mathcal{D}$).

Problem 2: Decision Trees, Overfitting

(5+10+10+5+10+10=50 points)

(a)

Table 3 reports the count for the diagnosis labels of the training set.

Label	*Count*
1	484
0	416

Table 3: Dataset $\mathcal{D}$ with Predictions

While split pretty evenly, the majority label is 1. Therefore, if use 1 as our predition for every case we get the accuracy scores reported in Table 4.

Set	*Accuracy score*
Training	53.78%
Test	50.60%

Table 4: Accuracy scores for baseline model

(b)

Table 5 reports the accuracy scores for the implementation of the decision tree with default parameters.

Set	*Accuracy score*
Training	100%
Test	56.18%

Table 5: Accuracy scores for decision tree

As we can see, we get perfect prediction on the training set while only slightly better prediction accuracy, i.e. 56.18% > 50.60%.

While we get a accuracy of 100% on the training, this might not be the optimal model has we might be overfitting our training set which could yield poor predictive power. This is the classic trade-off between variance and bias.

(c)

Looking at the attributes of our fitted `DecisionTreeClassifier`, we have a `max_depth` of 21.

Figure 1 show the accuracy scores for different values of `max_depth`.

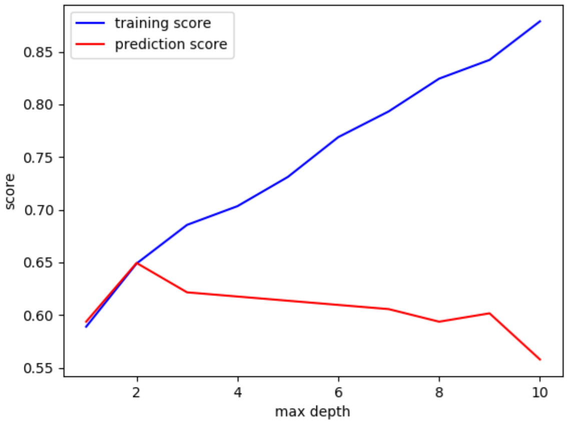

Figure 1: Accuracy scores for decision tree

As we can see, while the accuracy score on the training increases steadily as `max_depth` increases, the prediction score increase up to `max_depth`=2 and then starts decreasing. This indicates that we might be better off with a simpler model than one with `max_depth`=21.

(d)

Akin to p-hacking, if we play around with the parameters enough, we will find a spurious model that will seem to be able to predict the data perfectely solely for this peculiar context. For robustness, it is better to have a simpler model.

(e)

Figure 2 show the accuracy scores for the 4-fold cross-validation for different values of `max_depth`.

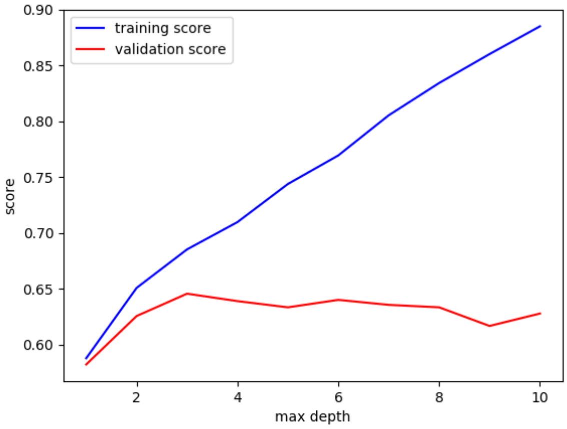

Figure 2: 4-fold validation curves

The depth that gives us the best perfomance is 3. We therefore use `max_depth`=3 and report the accuracy scores on the whole training set and test set in Table 6.

Set	*Accuracy score*
Training	68.56%
Test	62.15%

Table 6: Accuracy scores with `max_depth` = 3

(f)

Figure 3 show the accuracy score for the 4-fold cross-validation for different number of k-best features and `max_depth` = 10.

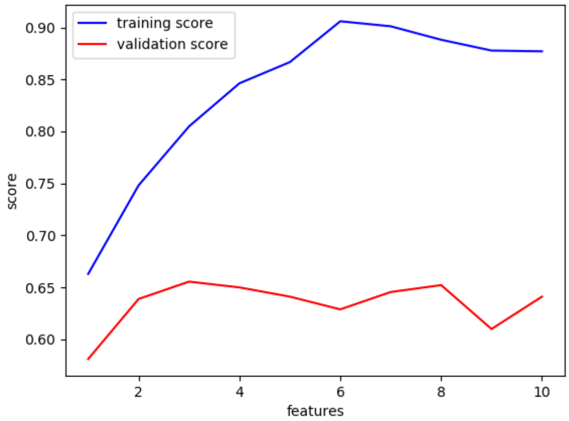

Figure 3: 4-fold validation curves with `max_depth` = 10

It is not clear from just the graph but it seems like the best number of k-best features is 3.
Table 7 reports the accuracy score for using the first 3-best features and a maximum depth of 10.

Set	*Accuracy scores*
Training	79.67%
Test	68.13%

Table 7: Accuracy score with `max_depth` = 10 and 3-best features

Therefore, this simpler decision tree seems to outperform every single model we considered so far.

Problem 3: Hypothesis Testing (15+10=25 points)

(a)

First, Table 8 reports the contingency table for the perfomance of m_a and m_b.

$m_a \backslash m_b$	Correct	Wrong	Total
Correct	75	10	85
Wrong	8	2	10
Total	83	12	95

Table 8: Contingency table

Clearly, we have a sample prediction error for m_a of $\hat{p}_a = 1 - \frac{85}{95} = 10.526\%$ and for m_b of $\hat{p}_b = 1 - \frac{83}{95} = 12.632\%$.

Before computing the confidence interval, we have to check if we approximate the distribution of prediction errors with a normal. To do this, we need to verify that $np(1-p) > 5$. Since $n\hat{p}_a(1-\hat{p}_a) = 8.947 > 5$ and $n\hat{p}_b(1-\hat{p}_b) = 10.484 > 5$, we are in the clear.

As shown in class, the 95% confidence interval when the normal distribution assumption is justified is given by the following formula:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hence, we have $10.526\% \pm 6.171\%$ for m_a and $12.632\% \pm 6.680\%$ for m_b.

(b)

For the McNemar's test, we want to test $H_0 : p_a = p_b$ versus $H_1 : p_a \neq p_b$. But before computing the statistic, it is useful to relabel the contingency table as in Table 9.

$m_a \backslash m_b$	Correct	Wrong
Correct	a	b
Wrong	c	d

Table 9: Contingency table

If both models are similarly good, they should not be able to outperfom the other by classifying correctly more frequently when the other is incorrect. Hence, the McNemar's test can computed the following way:

$$T = \frac{(b-c)^2}{b+c}$$

In our case, we have

$$T = \frac{(10-8)^2}{10+8} = \frac{2}{9} \approx 0.222$$

Under certain assumptions, $T \sim \chi^2_1$. Since the 95% critical value for χ^2_1 is 3.84, we would not reject the null. Moreover, using the inverse of the CDF of χ^2_1 we get that the p-value of T is equal to 0.637 well over 0.05.

Now, the previous result hold only if certain assumptions are met. One of them is that the data points are independent of each other. This seems reasonables in the case of cancerous nodes.

Another assumption is that $b+c$ is large enough for $T \sim \chi^2_1$. The rule of thumb is $b + c \geq 25$ which is sadly not met in our example.

Thankfully, we can use the binomial distribution with $n = b + c$ and $p = \frac{1}{2}$ to estimate if discrepancy between between b and c is likely. In fact, let x be the measure of Correct \Wrong and y be the measure of Wrong \Correct , we want to know the probability that we get b (c) or more extreme results, i.e. $P(x \geq b \mid n = b + c, p = .5) + P(y \leq c \mid n = b + c, p = .5)$.

This yields the following derivation of the p-value:

$$\begin{aligned}\text{p-value} &= P(x \geq 10 \mid n = 18, p = .5) + P(x \leq 8 \mid n = 18, p = .5)\\ &= 1 - P(x = 9 \mid n = 18, p = .5)\\ &= 0.815\end{aligned}$$

Thus, we can't reject the hypothesis.