Q5. You are given the following data set (x,y) (where $y=f(x)+\sigma^2$): (1,0.9), (2,2.2), (3,3), (4,4.1), (5,4.8), (6,6). You trained your ML algorithm on this dataset to obtain a model with $\hat{f}(x)=\theta_1 x+\theta_2 x^2$ where the identified values were $\theta_1=1$ and $\theta_2=0.01$.

a. Please compute the Residual Sum of Squares (RSS) for the given data set where $RSS=\sum_{i=1}^{N}(\hat{f}(x_i)-y_i)^2$. Please show the steps involved. [Note that MSE = RSS/N where N is the number of data points.] [3]

$$RSS = \sum_{i=1}^{6}(\hat{f}(x_i)-y_i)^2 = (1.01-0.9)^2 + (2.04-2.2)^2 + (3.09-3)^2$$
$$+ (4.16-4.1)^2 + (5.25-4.8)^2$$
$$+ (6.36-6)^2$$
$$= \cancel{0.04} + 0.012 + 0.026 + 0.01$$
$$+ 0.004 + 0.202 + 0.129$$
$$= 0.383$$

b. Your goal is to minimize RSS. If you know the ground truth that values in decimal are due to noise component - please design a regularization function to reduce the effect of x^2 i.e., present the new RSS function you plan to minimize. Please explain in detail how this would achieve the desired effect. [3]

(B∨D) → (L∧M)

Q1. Tick *all* correct answers; Cross *all* wrong answers. Negative marks of 25% will be awarded for each wrongly marked answer.

(i) If Ram took the bus (B) or drove in his own car (D), then he arrived late (L) and missed the first session (M).

(a) ¬B∨¬D∨L∨M

(b) ¬B∧¬D∨L∨M

(c) (B∨D) → (L∧M)

(d) B∨D∨L∨M [4]

(ii) If the king does not castle (K=King castles) and the pawn advances (P) then either the bishop is blocked or the rook is pinned (R).

(a) (¬K∧P) → (B∨R)

(b) K∨P∨B∨R

(c) ¬K∨¬P∨B∨R

(d) K∨¬P∨B∨R [4]

Q2. Prove the resolution rule ($P \vee Q, \neg Q \vee R \Rightarrow P \vee R$) using truth-tables. [10]

P	Q	R	P ∨ Q	¬Q∨R	P∨R	[(P∨Q)∧(¬Q∨R)] → (P∨R)
T	T	T	T	T	T	T
T	T	F	T	F	T	~~F~~ T
T	F	T	T	T	T	T
T	F	F	T	T	T	T
F	T	T	T	T	T	T
F	T	F	T	F	F	~~F~~ T
F	F	T	F	T	T	T
F	F	F	F	T	F	T

Q3. Prove using resolution: $P \rightarrow \neg Q, \neg Q \rightarrow R \Rightarrow P \rightarrow R$. (Represent a clause with its number.) [10]

Clauses being resolved	Substitution required	New resulting clause	Clause number (of new clause)

c. One issue with real data sets is you would not know the effect of noise and hence you would not know that the decimal values are due to noise. Given this you would not be able to design a regularization function that would act on the specific terms.
Two popular general regularization functions in literature are Ridge and Lasso regularization. Ridge regularization also called an L2 penalty, is going to square your coefficients. Lasso regularization or an L1 penalty, is going to take the absolute value of your coefficients.

RSS with Ridge regularization: $RSS=\sum_{i=1}^{n}\left(\hat{f}(x_i)-y_i\right)^2+\alpha\sum_{j=1}^{n}(\theta_j)^2$

RSS with Lasso regularization: $RSS=\sum_{i=1}^{n}\left(\hat{f}(x_i)-y_i\right)^2+\alpha\sum_{j=1}^{n}|\theta_j|$

Assuming $\alpha=1$, which of the two regularization functions would be better for your $\hat{f}(x)$? Please explain your choice with mathematical reasoning. [4]

Q4. Suppose a training set consists of points $x_1, x_2, \ldots, x_n$ and real values y_i associated with each point x_i. We assume there is a function with noise $y = f(x) + \varepsilon$, where the noise ε has zero mean and variance σ^2. Please provide all steps of derivation for

$$E\left[\left(y-\hat{f}(x)\right)^2\right] = \left(Bias\left[\hat{f}(x)\right]\right)^2 + Var\left[\hat{f}(x)\right] + \sigma^2$$

where $\hat{f}(x)$ is the best approximation for $f(x)$ identified by the machine learning algorithm. [8]

$$E[(y-\hat{f}(x))^2] = E[(f+\varepsilon-\hat{f})^2]$$

$$= E[(f+\varepsilon-\hat{f}+E(\hat{f})-E(\hat{f}))^2]$$

$$= E[((f-E(\hat{f}))+\epsilon+(E(\hat{f})-\hat{f}))^2]$$

$$= E[(f-E(\hat{f}))^2] + E[\epsilon^2] + E[(E(\hat{f})-\hat{f})^2]$$

$$+2E[(f-E(\hat{f}))]E(\epsilon) + 2E[\epsilon]E[(E(\hat{f})-\hat{f})]$$

$$+2E[(f-E(\hat{f}))]E[(\hat{f}-E(\hat{f}))]$$

$$= E[f^2] + E[E[\hat{f}]^2] - 2E[f]E[E(\hat{f})]$$

$$+ E[\hat{f}^2] + E[E[\hat{f}]^2] - 2E[\hat{f}]E[E[(\hat{f})]]$$

$$\cancel{+2E[f]E[E\hat{f}}$$

$$+2E[f]E[\epsilon] - \cancel{2E[\epsilon]E[E[\hat{f}]]}$$

$$+2E[\cancel{E[\hat{f}]}]E[\epsilon] - 2E[\hat{f}]E[\epsilon]$$

$$+2E[f]E[\hat{f}] - 2E[f]E[E(\hat{f})] - E[\hat{f}]^2$$

12

Q6. Please answer the following questions:

a. What is the AlphaGo system? Please describe idea behind the system in 2-3 sentences ? You should mention atleast three specific ideas or techniques used as part of this system. [1 + 2]

The AlphaGo is a Machine Learning system trained on the Monte Carlo tree method to play chess. Within 4 hours of training it beat the reigning chess robot champion.

b. What does dimensionality reduction mean ? Please explain how it can be used to reduce overfitting ? [1 + 2]

Dimensionality reduction deals with reducing the features used to train the model. This is done because disregarding features not relevant to the reasons for training the model can better help create a generalised model which doesn't memorize the ~~data~~ training set.

c. What is feature engineering ? Please explain how binning can be used to perform feature engineering ? [1 + 2]

Feature engineering is the process of creating new features to train your model based on pre-existing ones.

d. What is the goal of Power transformation ? Please present a diagram with some description that conveys the essence of your explanation ? [1.5+1.5]

Transformations are a way of making your features more easily possible as an input to your model. Power transformations involve modifying the power of your features to achieve the same.

Q7. Given the following frequent itemsets what candidates will Apriori compute for the next database scan? Show your steps.

(i) AB, AC, AD, BC, BD, CD, AE [6]

i.) We first create a set of sets of three by matching the given sets so that they have at most one subitem different. Then, we have:

[ABC, ABD, ABE, ACD, ACE, ADE, BCD]

ii.) From this set, we eliminate itemsets who do not have subsets in the given set. So, we have,

[ABC, ABD, ACD, BCD]

This is our final answer.

(ii) ABC, ABD, ACD, BCD, BCE, CDE [6]

i.) Similar to above, the first step produces the following:

[ABCD, ABCE, ACDE, BCDE]

ii.) Now, eliminating results in our final answer:

[ABCD]