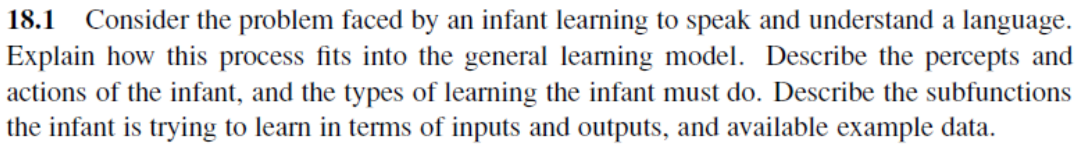
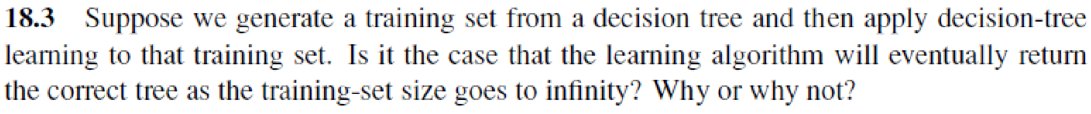
Q1.



18.1 Consider the problem faced by an infant learning to speak and understand a language. Explain how this process fits into the general learning model. Describe the percepts and actions of the infant, and the types of learning the infant must do. Describe the subfunctions the infant is trying to learn in terms of inputs and outputs, and available example data.

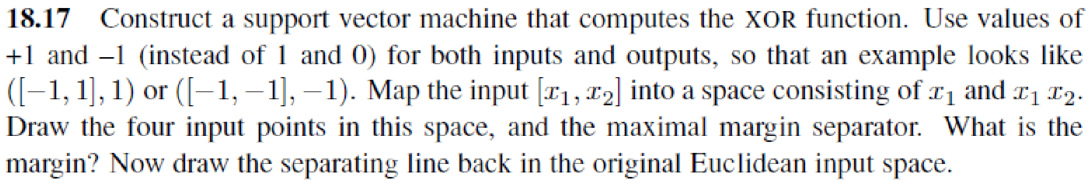
The simulation model of learning environment only considers the exchange of speech. In fact, the physical context of each utterance is crucial: Children must see the context in which "watermelon" is said before they can learn to associate "watermelon" with watermelon. Therefore, the environment consists not only of other people, but also of the physical objects and events where the conversation takes place. Auditory sensors detect speech, while other senses (primarily vision) provide information about the physical environment. The effectors are the articulatory organs and motor abilities that allow the infant to respond to speech or to elicit verbal feedback.

Performance criteria can be simply a general utility function for infants, so infants perform reinforcement learning to perform and respond to speech behaviors to improve their well-being. However, humans' built-in mimicry ability suggests that producing sounds similar to those made by others is a goal in itself. Children are also exposed to examples of supervised learning: adults say the name of the appropriate object when indicating it. Thus, adult-generated sentences provide positive examples of labeling, and adult responses to infant speech acts provide further categorical feedback.



18.3 Suppose we generate a training set from a decision tree and then apply decision-tree learning to that training set. Is it the case that the learning algorithm will eventually return the correct tree as the training-set size goes to infinity? Why or why not?

This algorithm will return the logically equivalent tree, any two decision trees defined on the same set of attributes that are consistent for all possible examples. The actual form of the tree may vary, because there are many different ways to represent the same function. The root attribute of the original tree may not actually be the one the information will select



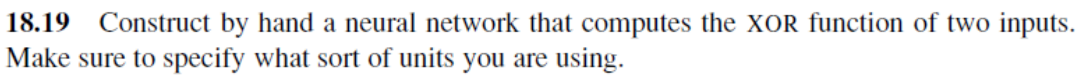
18.17 Construct a support vector machine that computes the XOR function. Use values of +1 and –1 (instead of 1 and 0) for both inputs and outputs, so that an example looks like $([-1,1],1)$ or $([-1,-1],-1)$. Map the input $[x_1, x_2]$ into a space consisting of x_1 and $x_1\,x_2$. Draw the four input points in this space, and the maximal margin separator. What is the margin? Now draw the separating line back in the original Euclidean input space.

The examples map from [x1, x2] to [x1, x1, x2] coordinates are as follows:

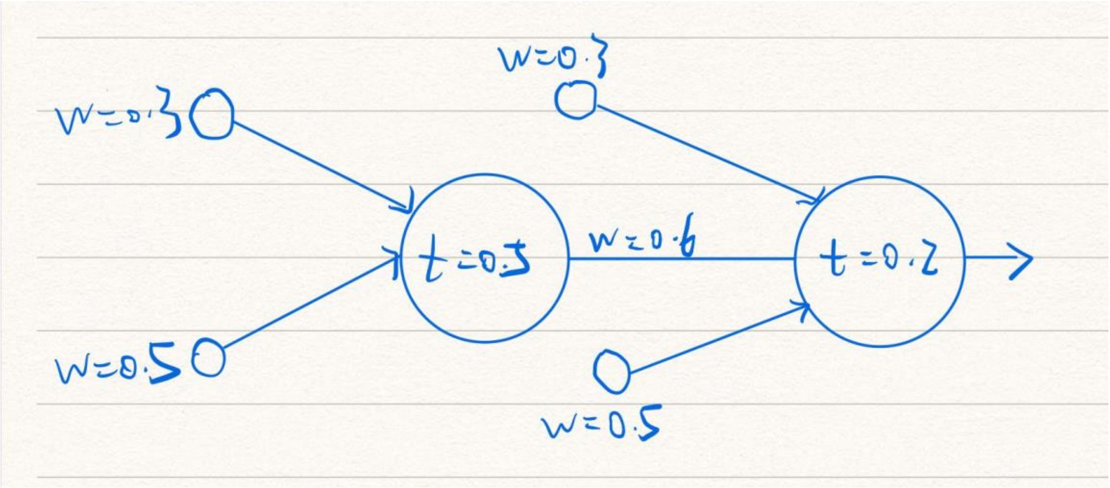
[−1,−1] (-) maps to [−1, +1], [−1, +1] (+) maps to [−1,−1], [+1,−1] (-) maps to [+1,−1],

[+1, +1] (+) maps to [+1, +1]

The maximum margin separator is the line x1*x2 =0 with a margin of 1. The separator corresponds to the x1 =0 and x2 =0 axes in the original space, and this can be thought of as the limitation of a hyperbolic separator with two branches.

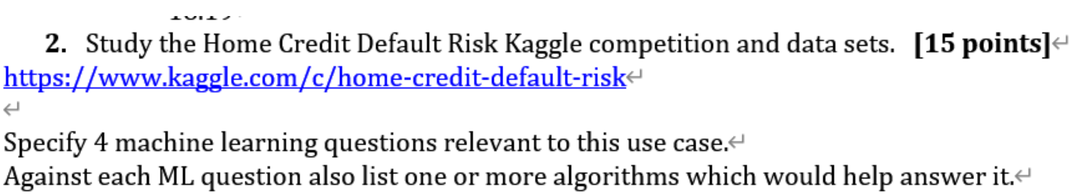


18.19 Construct by hand a neural network that computes the XOR function of two inputs. Make sure to specify what sort of units you are using.



The hidden layer computes AND, while the output layer computes OR but weights the output of the hidden node negatively.

Q2

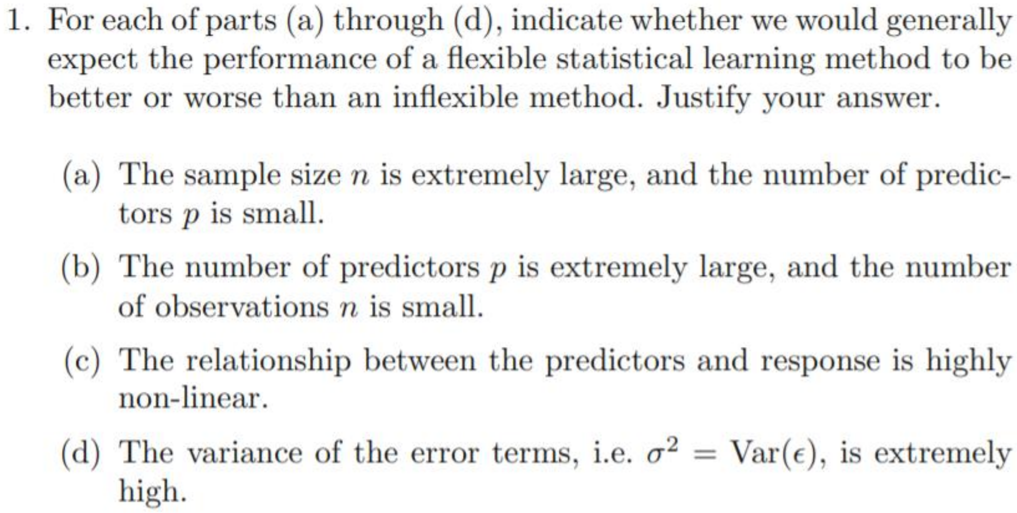


2. Study the Home Credit Default Risk Kaggle competition and data sets. **[15 points]**
https://www.kaggle.com/c/home-credit-default-risk

Specify 4 machine learning questions relevant to this use case.
Against each ML question also list one or more algorithms which would help answer it.

1. What the connection between monthly balance and repayment?
   Algorithm: Linear programming
2. How to group the users by their credit?
   Algorithm: SVM
3. How to judge whether the users could gain the loan?
   Algorithm: Bayes net
4. Using other data as the probability, how to judge the probability of the repayment?
   Algorithm: Decision trees

Q3



1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The relationship between the predictors and response is highly non-linear.

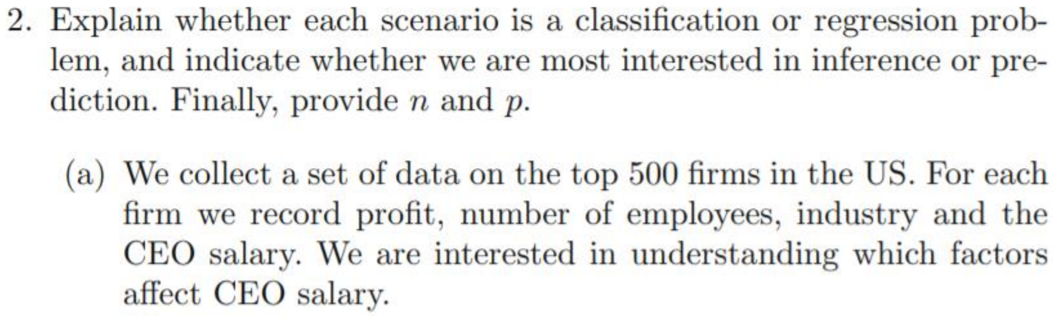
   (d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

(a) We could use a large number of n and flexible learning method to try to avoid overfitting.

(b) If p is large and n is small. Then there is possible that flexible learning method would lead to overfit. So we use inflexible method.

(c) A highly non-linear method should be applied a flexible learning method.

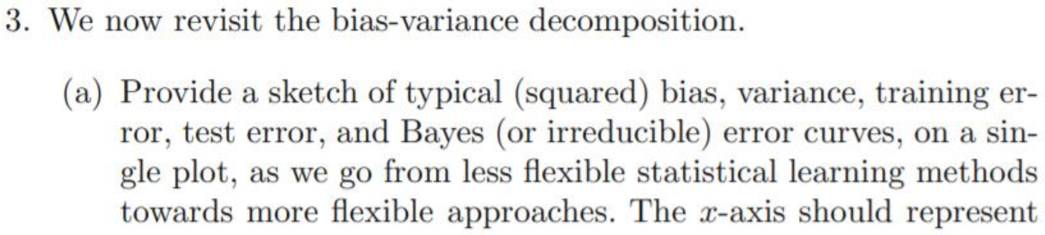
(d) The variance is very large, so we use inflexible method to avoid overfitting.



2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

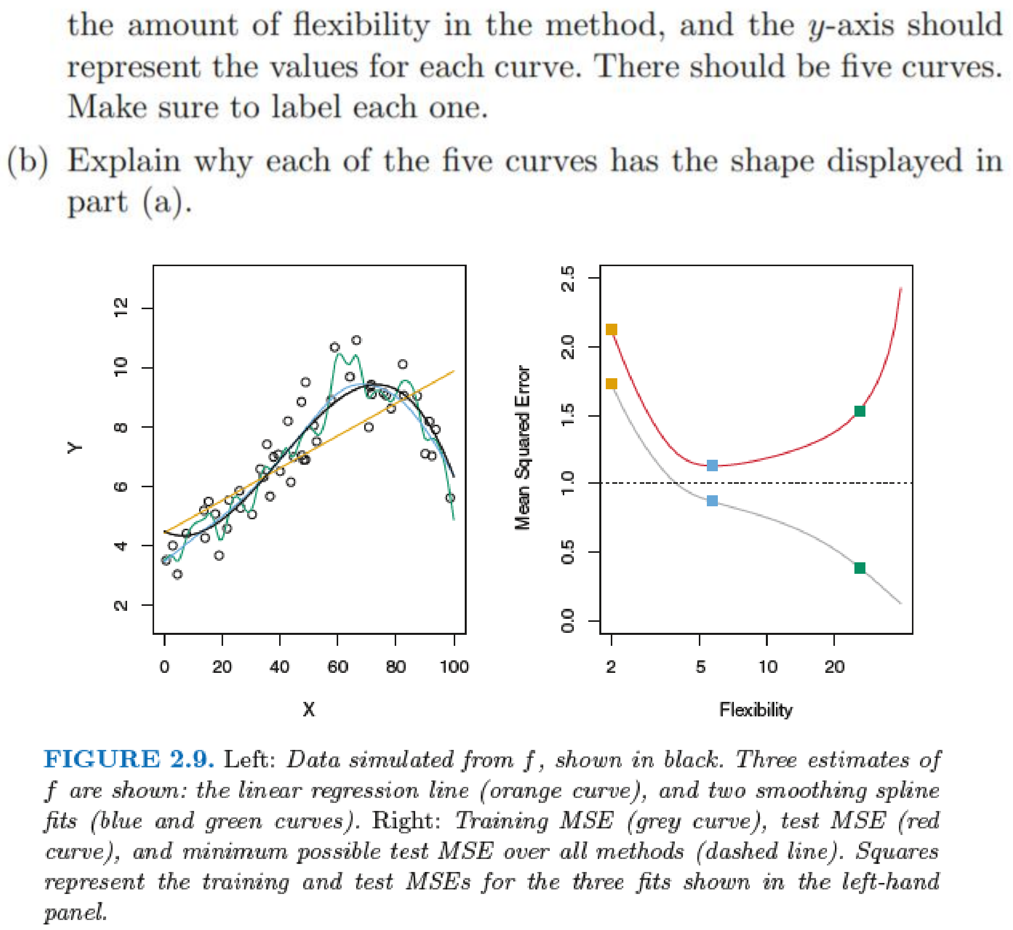
   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

We are most interested in regression problem, so the n=500 and p=3



3. We now revisit the bias-variance decomposition.

   (a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent



the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(b) Explain why each of the five curves has the shape displayed in part (a).

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

(a)

(b) As the flexibility of the learning method increases, training errors show a steady decline.

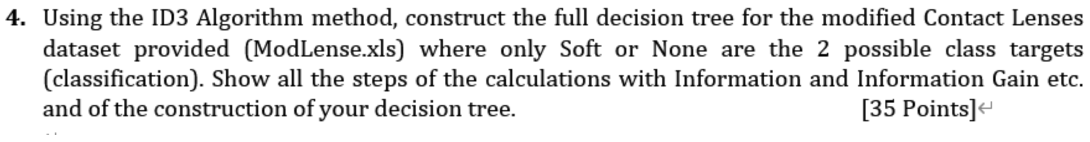
The test error is the red curve, which initially decreases as the flexibility increases, but begins to increase again as the flexibility continues to increase.

The Bayesian irreducible error is a constant dashed line. Note that the point at which the test error is closest to the Bayesian error will be the optimal operating point for the system.

The distance between the lowest point on the test error curve

The Bayesian error indicates how much deviation exists in a given learning process,

Q4



4. Using the ID3 Algorithm method, construct the full decision tree for the modified Contact Lenses dataset provided (ModLense.xls) where only Soft or None are the 2 possible class targets (classification). Show all the steps of the calculations with Information and Information Gain etc. and of the construction of your decision tree. [35 Points]↵

In ID3:

The IG for feature Age： 0.075

The IG for feature Astigmatism： 0.001

The IG for feature Tear-production： 0.344

So the optimal feature is Tear-production

Then:

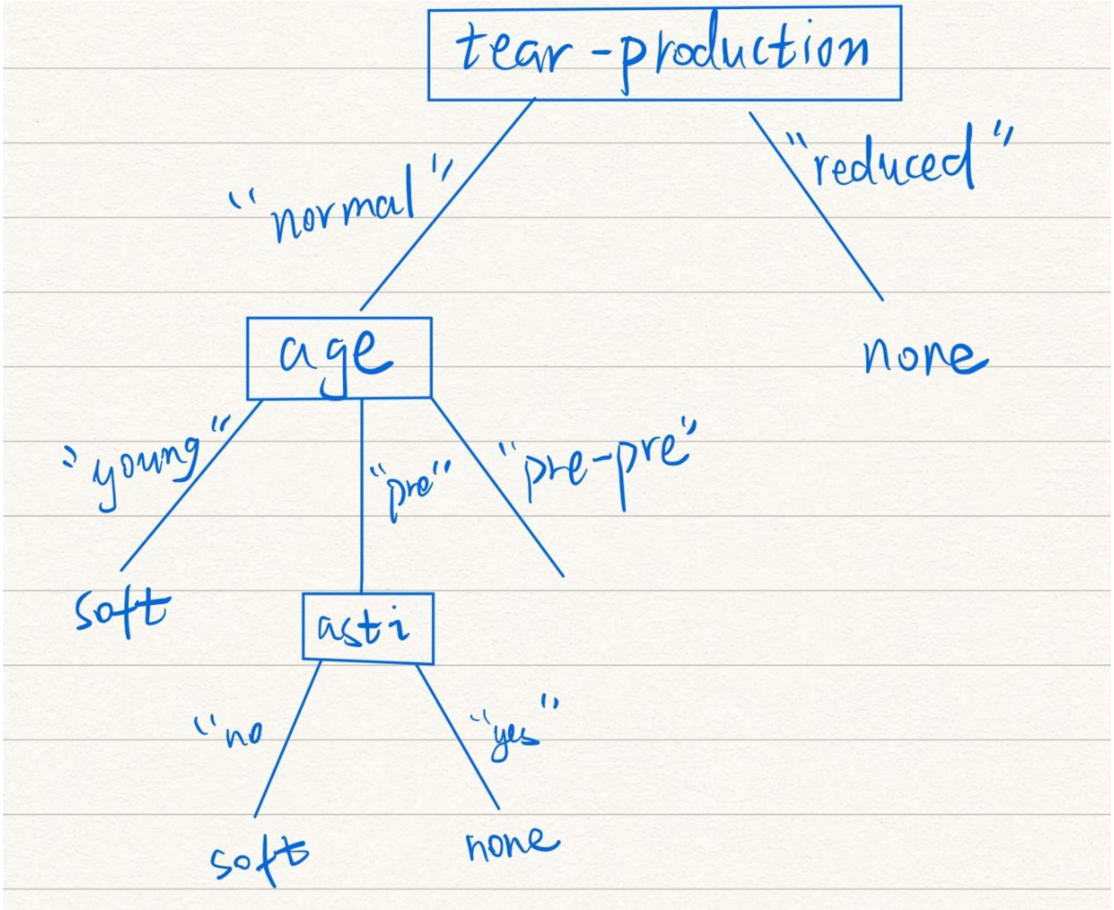
The IG for feature Age：0.204

The IG for feature Astigmatism：0.003

The optimal feature is Age

The for feature Asti, we need to judge the result based on the majority of samples.

The tree is like:


tear-production
"normal"
"reduced"
age
none
"young"
"pre"
"pre-pre"
soft
asti
"no
"yes"
soft
none