Data Mining

Homework 2

Important Notes:

1. Submit in electronic form before 11:59pm on Wednesday, April 19, 2017
2. No late homework will be accepted.
3. The homework should be completed and submitted by each individual.
4. The homework should be submitted through Gradescope. Entry Code: **9BW66M**
5. The homework should be written in English.
6. The HW is worth it 10 points.
7. The [Research] questions require from you to do some research on the Web and get to understand things that were not covered during the lecture.
8. For questions, please use Piazza (English only!)

Linear Regression [10 pts]

Assume that you wish to predict the price of a car given the following set of eleven features/attributes; the values of these features can be found at the second column. Further assume that you consider a linear combination of these features as the model (hypothesis) to predict a car's price.

Feature/Attribute	Values
1. make	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda,mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
2. num-of-doors	four, two
3. body-style	hardtop, wagon, sedan, hatchback, convertible.
4. fuel-type	diesel, gas
5. length	continuous from 141.1 to 208.1
6. width	continuous from 60.3 to 72.3
7. engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor
8. num-of-cylinders	eight, five, four, six, three, twelve, two
9. engine-size	Continuous from 61 to 326
10. horsepower	continuous from 48 to 288
11. peak-rpm	continuous from 4150 to 6600

1. A number of features above are nominal features (i.e. categorical). Consider for instance the make, the fuel-type, and the num-of-doors. How would you represent them in your hypothesis (one-hot representation, numerical value, or something else)? Explain your choice.

Make: numerical values, since people also pay for branding. Therefore I would give bigger brands a higher numerical value in comparison to other brands. The hard part would be the ranking of the brands.

Fuel-type: one-hot representation, because we can not say in advance that diesel is greater than gas.

num-of-doors: one-hot representation. Intuitive I'd say that more doors would increase the production cost, however some of the most expensive cars only have two doors. Therefore I couldn't say one value of num-of-doors would be greater than the other value and thus I'd use one-hot.

2. During the lecture I made the following statement: "*The penalty term in regularization is not fair if features are not on the same scale*". Explain why is this statement true (accompany the explanation with an example).

When two features are not on the same scale they'll have different contributions to the penalized terms. This is due to the fact that the penalized term is the sum of the squares of all coefficient, therefore the penalty term will not be fair.

E.g. when you have an independent variable "height". The height of human (Europe) is measured in meters, but another feature might display the height of some object on another scale (like millimeters). In this case the penalty term will obviously be unfair.

3. Assume that you construct a model (i.e. a hypothesis) using only three of the features in the table above: engine-size, horsepower, and fuel-type.
    a. Write the hypothesis/model.

$H\Theta(x) = \Theta_0 + \Theta_1$*(engine-size) + Θ_2*(horsepower) + Θ_3*(fuel-type)

With $H\Theta(x)$ being the carprice, and Θ_i the weight of the corresponding feature.

b. Develop a 2-degree polynomial of the three features including their interaction. Write the new model/hypothesis.

H = Θ_0 + Θ_1*(engine-size) + Θ_2*(horsepower) + Θ_3*(fuel-type) + Θ_4*(engine-size)^2 + Θ_5*(engine-size * horsepower) + Θ_6*(engine-size * fuel-type) + Θ_7*(horse-power)^2 + Θ_8*(horse-power * fuel-type) + Θ_9*(fuel-type)^2

With HΘ(x) being the carprice, and Θ_i the weight of a feature.

4. Why is it necessary to use a test set, which is different from the training set to evaluate the performance of a machine learning algorithm?

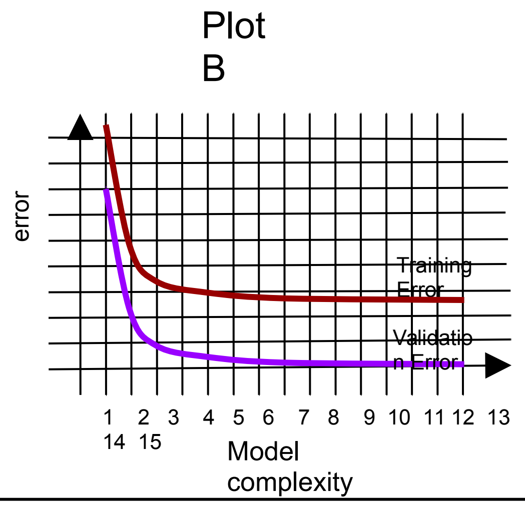
The training set is used to train the model. When the model is "trained" we can evaluate the performance and see how well our model is fitted by testing the model with features (with known targets) and compare the result of our model and the "real value".

When you would test your model using the same training set and get 100% accuracy this would not mean your model is 100% accurate (due to overfitting, noise etc.). However when you achieve 100% on a distinct set (the test set) this will give a better indicator of the ability of the model.

5. When you construct more than one hypothesis/model, and you wish to pick the best performing one (i.e. model selection), why not use the test set for that, but instead construct a validation set?

The model's parameters are adjusted to the best accuracy based on the test set. Because the test data is used to adjust these parameters it can no longer be used to evaluate the correctness of our model. Therefore we need a distinct set.

6. Is the following graph a reasonable graph to observe during training? Explain your answer.



Intuitively the training error (used to make the model) should be lower than the validation error (used to adjust the model). The model is fitted to the training set, so it would be a weird outcome if the validation set yields smaller errors. However the graph can be reasonable when e.g. the training set consisted of many difficult cases to learn in contrast to the validation set, which could have consisted of many easy cases to learn.

7. Why you should not fix the test set but instead do a K-fold cross-validation?

When using K-fold you are using all data to generate your training- and test-set. Each k-subset will only be used once to validate, and therefore the average error will be much more correct in comparison to having a fixed test set. This is due to the fact that the results of the fixed test set could for example consist of only samples of 1 class.

8. [Research] Can you think of the advantages and the disadvantages of a large K, compared to a small K in K-fold cross-validation?

Advantages: when having a large K you reduce the probability of a set containing only samples of one class.

Disadvantages: takes more time to evaluate.

9. [Research] The Lasso regularization can lead to feature selection (i.e. it makes some of the parameters theta become zero). This is often not the case for the Ridge regularization. Can you explain why not?

In the ridge regularization the coefficients of the linear transformation are normal distributed in contrast to lasso, where the coefficient are Laplace distributed.