Introduction to Machine Learning
Homework 2: Model Order Selection*

Prof. Linda Sellie

1. If your training data set D consists of $N = 100$ points, $(\mathbf{x}^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$ where $\mathbf{x}^{(i)}$ consists of a single feature (i.e. $\mathbf{x}^{(i)} = [x_1^{(i)}]$), and we fit two linear regression models

   model 1: $w_0 + w_1x_1 + w_2x_1^2$.

   model 2: $w_0 + w_1x_1$

   - When using model 1, what transformation function[1] $\Phi(\mathbf{x})$ would we use? Your answer should have the form $\Phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \phi_1(\mathbf{x}) \\ \vdots \\ \phi_{\tilde{d}}(\mathbf{x}) \end{bmatrix}$ (Include x_0 as part of your transformation.)
   - For model 1, express RSS (residual sum of squares) in terms of $\Phi(\mathbf{x})$.
   - Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1x_1 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}? Which model would we expect to have a smaller generalization error, E_{out}? Explain.
   - Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}? Which model would we expect to have a smaller generalization error, E_{out}? Explain.

2. A medical researcher wishes to evaluate a new diagnostic test for cancer. A clinical trial is conducted where the diagnostic measurement y of each patient is recorded along with attributes of a sample of cancerous tissue from the patient. Three possible models are considered for the diagnostic measurement:

   - Model 1: The diagnostic measurement y depends linearly only on the cancer volume.
   - Model 2: The diagnostic measurement y depends linearly on the cancer volume and the patient's age.
   - Model 3: The diagnostic measurement y depends linearly on the cancer volume and the patient's age, but the dependence (slope) on the cancer volume is different for two types of cancer – Type I and II. (Hint: Use a variable x_3, which is assigned the value 1 if the cancer is Type I, and x_3 has the value 0 if the cancer is Type II.)

*Some of these questions are adapted from Prof. Rangan's homework.

[1]We learned that by performing a nonlinear feature transformation, we could fit a non-linear model as if it is a linear model.

(a) Define variables for the cancer volume, age, and cancer type and write a linear model for the predicted value $\hat{y}$ in terms of these variables for models 1 & 2 above.

(b) What is the number of parameters in models 1 & 2? Which model is the most complex?

(c) Since the models in part (a) is linear, given training data, we should have $\hat{\mathbf{y}} = X\mathbf{w}$ where $\hat{\mathbf{y}}$ is the vector of predicted values on the training data, X is a design matrix (feature matrix), and $\mathbf{w}$ is the vector of parameters. To test the different models, data is collected from 100 patients. The records of the first three patients are shown below:

Patient ID	Measurement y	Cancer type	Cancer volume	Patient age
12	5	I	0.7	55
34	10	II	1.3	65
23	15	II	1.6	70
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

For model 1 in part (a), based on this data, what are the first three rows of the matrix X?

For model 2 in part (a), based on this data, what are the first three rows of the matrix X?

(d) To evaluate the models, 10-fold cross-validation is used with the following results.

Model	training MSE	test MSE
1	2.0	2.01
2	0.7	0.72
3	0.65	0.74

Which model should be selected?

3. Suppose you trained your data[2] on three different models and then plotted how well the different fitted models performed with varying amounts of data:

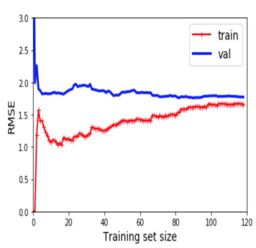

Figure 1: A

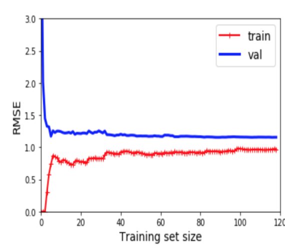

Figure 2: B

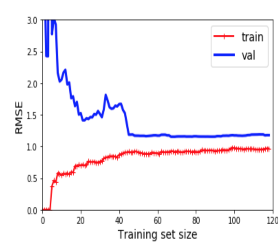

Figure 3: C

What can you say about overfitting and underfitting? What can you say about the number of examples and the model's fit?

[2]The data remains the same.

4. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach[3] for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred ?[4]

5. Consider a binary classification problem ($y \in \{0, 1\}$), where the iid examples

$$D = (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})$$

are divided into two disjoint sets D_{train} and D_{val}.

- Suppose you fit a model h using the training set, D_{train} and then estimate its error using the validation set, D_{val}. If the size of D_{val} was 100 (i.e. $|D_{val}| = 100$), how confident are you the true error of h is within 0.1 of its average error on D_{val}?
- Repeat the previous question where now $|D_{val}| = 200$ (i.e you have 200 examples in your validation set).
- (Do not turn in this question) When dividing the set of examples D into two sets, how large should you make D_{val} if you wanted to be 90% confident that the true error of h is within 0.05 of the average error your hypothesis makes on D_{val}.

6. Suppose you are given the following dataset, where the target variable is MED:

RM	RAD	DIS	MED
6.6	1	4.0	24.0
6.4	2	5.0	21.6
7.2	2	5.0	34.7
6.4	2	5.0	21.6
7.2	2	5.0	34.7

Using the data above, write the *equation* derived in the lecture notes to compute the closed form solution for ridge regression where $\lambda = 0.1$. You do not need to calculate the coefficient vector - just set up the formula using the numbers given above.

7. Write the gradient descent algorithm (vectorized or not) for ridge regression

8. (Do not turn in this question) For the training examples in question 3 from homework assignment 1, write the closed-form solution for ridge regression when $\lambda = 0.1$.

9. (Do not turn in this question) For each of parts 9a through 9d, indicate whether we would generally expect the performance of a flexible (complex) hypothesis class (aka complex model class) to be better or worse than an inflexible (simple) hypothesis class (aka simple model)..[5] Justify your answer.

   (a) The sample size N is huge, and the number of features d is small.

   (b) The number of features d is huge, and the number of observations N is small.

[3] i.e., more complex hypothesis class (model class) or less complex hypothesis class (model class).
[4] This question is a modified version of a question in ISLR.
[5] This question is a modified version of a question in ISLR.

(c) The relationship between the features and labels is highly non-linear.

(d) The variance of the noise, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

10. (Do not turn in this question) Bias-variance decomposition

    - Provide a sketch of typical (squared) bias, variance, training error, and test error on a single plot as we go from less flexible statistical learning methods toward 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 four curves. Make sure to label each one.[6]
    - Explain why each of the curves has the shape displayed in part (a).

11. (Do not turn in this question) Given a dataset of N items, how would you use k-fold cross-validation to decide which hypothesis class (model) to use if your choices were

    - fitting a linear regression model on the data or
    - fitting a polynomial of degree 2 model on the data

[6]This question is a modified version of a question in ISLR. Please also note that the lecture notes show a plot of 3 of these curves.