Introduction to Machine Learning
Homework 3: Model Order Selection*

Prof. Linda Sellie

Spring 2018

1. For each of the following pairs of true functions $f(\mathbf{x})$ and model classes $h(\mathbf{x})$ determine: (i) if the model class is linear; (ii) if there is no under-fitting; and (iii) if there is no under-fitting, what is the true parameter?

   (a) $f(\mathbf{x}) = 1 + 2x$, and $h(x) = w_0 + w_1 x + w_2 x^2$

   (b) $f(\mathbf{x}) = 1 + x + 3x^2 + 4x^3$, and $h(x) = w_0 + w_1 x + w_2 x^2$.

   (c) $f(\mathbf{x}) = (x_1 - x_2)^2$ and $h(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$.

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 of Type II.)

   (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.

   Extra credit: Do the same for model 3. For Model 3, you will want to use one-hot coding as mentioned above.

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

*These questions are adapted from Prof. Rangan's homework.

(c) Since the models in part (a) are 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
⋮	⋮	⋮	⋮	⋮

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?

Extra credit: For model 3 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	0.70

Which model should be selected?

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

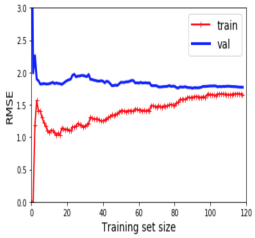

Figure 1: A

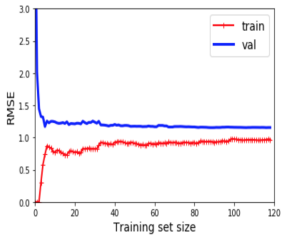

Figure 2: B

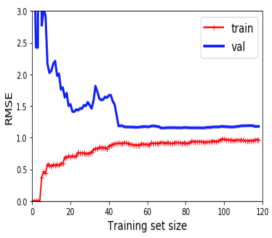

Figure 3: C

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