Artificial Intelligence

COMS 4701 Section 2 – Fall 2014
`http://www1.ccls.columbia.edu/~ansaf/4701/`

Home Work n°3: Machine Learning – Additional question (Optional)

Due Wednesday November 19th, 2014 @11:55pm

Preliminaries

This is an appendix to assignment 3. This problem is optional and will get you extra points if you do it. Consider this question as an optional part in Problem 1.

Problem 1: Regression

3. Regression with polynomial fitting (optional)

We are interested in studying the relationship between age and height (statures) in girls aged 2 to 20 years old. We think that this can be modeled with a polynomial regression model. Unfortunately, we only have a small sample to study. Each example has one feature *Age* along with a numerical label *Height*. We will use the dataset **girls_2_20_train.csv**. The linear model we would like to derive has the form:

$$\text{Height} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Age}^2 + \cdots \beta_d \times \text{Age}^d$$

where d is the degree of the polynomial.

1. Write a python code that uses the normal equation to finds the β's of the model. Experiment with various values of $d \in \{0, \cdots 5\}$ (Hint: Don't forget to add the intercept. Add new features that are power of the age to your data matrix).

2. Plot the data points along with the regression function obtained for each degree $d \in \{0, \cdots 5\}$. You should obtain 6 curves.

3. For what degree, do you observe under-fitting? over-fitting?

4. Use a validation set **girls_2_20_validation** to pick the best degree d. For this purpose, calculate the mean square error for training and validation for each d. Plot two curves one for the training error and one for the validation error. Use the same figure. Use this figure to infer the best d.