Exercises
Introduction to Machine Learning
FS 2018

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s18`
Email questions to:
Natalie Davidson, natalie.davidson@inf.ethz.ch

Series 2, Mar 16th, 2018 (Model selection and Classification)

Problem 1 (Model selection and cross-validation):

Suppose we are given a noise-free set of points $X = \{x_i\}_{i=1}^n \subset (-1, 1)$, $Y = \{\sin(x_i)\}$, which we want to fit with a polynomial, but we do not know which degree to choose. Suppose our candidate polynomial families are $\mathcal{P}_k = \{\mathbb{P}_{2i+1}\}_{i=0}^k$, where $\mathbb{P}_{2i+1}$ denotes the family of polynomials with real-valued coefficients of maximum degree $2i + 1$. We want to find the optimal hyperparameter value $\hat{k} \in \{1, \ldots, k\}$.

Given a family of polynomials $\mathbb{P}_{2\ell+1}$ and a training set, suppose we have an oracle (i.e. an exact algorithm) that is able to find the polynomial $\hat{p} \in \mathbb{P}_{2\ell+1}$ with optimal coefficients with respect to the square loss objective

$$\mathcal{L}(X, Y, p) = \sum_{i=1}^{n} (y_i - p(x_i))^2, \quad p \in \mathbb{P}_{2\ell+1}.$$

1. Show that when the optimization is performed on each family in $\mathcal{P}_k$, the lowest score is achieved when $\hat{p} \in \mathbb{P}_{2k+1} \setminus \mathbb{P}_{2k-1}$ (i.e., $\hat{p}$ will be of degree $2k + 1$).

   Answer:
   Note: We should consider polynomials of the following type, not the odd-ordered polynomials stated in the question previously

$$\begin{aligned}
&\mathcal{P}_1 = w_1 x \\
&\mathcal{P}_3 = w_1 x - w_2 x^3 \\
&\mathcal{P}_5 = w_1 x - w_2 x^3 + w_3 x^5 \\
&\mathcal{P}_7 = w_1 x - w_2 x^3 + w_3 x^5 - w_3 x^7 \\
&\ldots
\end{aligned}$$

   We also remember the taylor-series approximation of $sin(x)$

$$sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \ldots$$

   The loss can be calculated as:

$$
\begin{aligned}
\mathcal{L}(X, Y, p, k) &= \sum_{i=1}^{n} (y_i - p(x_i))^2, \quad p \in \mathbb{P}_{2\ell+1}. \\
&= \sum_{i=1}^{n} (sin(x_i) - p_{2k+1}(x_i))^2 \qquad \text{all terms in taylor series expansion will be eliminated up to 2k+1} \\
&= \sum_{i=1}^{n} \mathcal{O}(x_i^{2k+3})^2 \\
&= \sum_{i=1}^{n} \mathcal{O}(x_i^{4k+6}) \\
&> \sum_{i=1}^{n} \mathcal{O}(x_i^{4(k+1)+6}) \\
&= \sum_{i=1}^{n} \mathcal{O}(x_i^{4k+10})
\end{aligned}
$$

2. What potential issue with using cross-validation does this demonstrate?

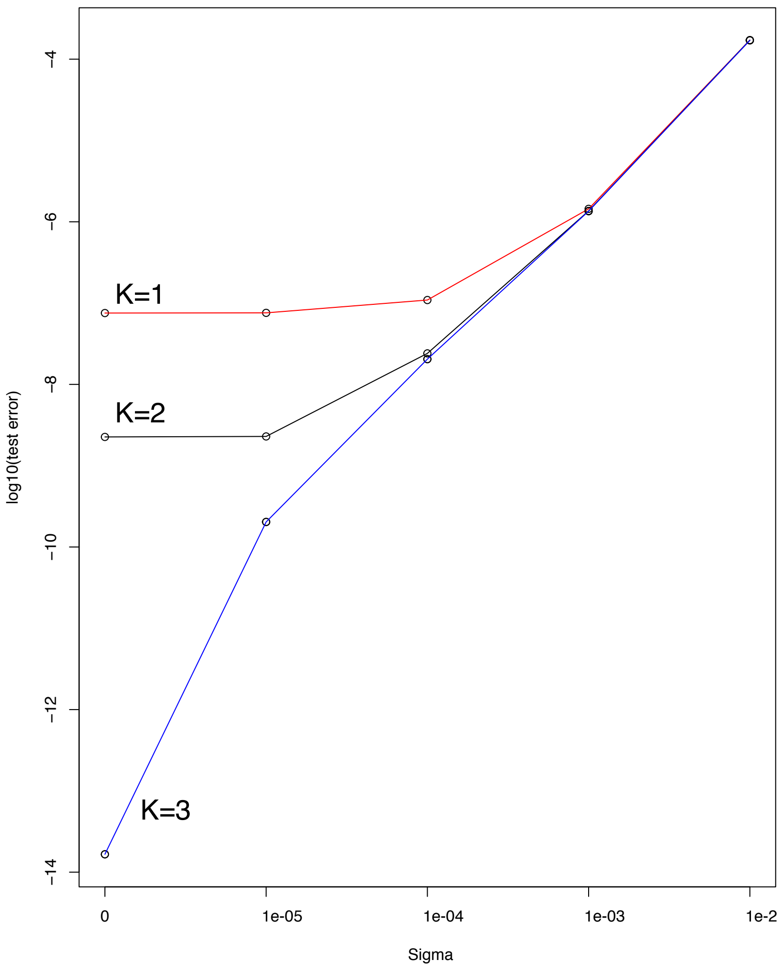
   Answer:
   When fitting a model, even when using cross-validation, we can still choose overly complex models. In this case, it occurred because of a special relationship between our data and the family of functions we were using to approximate it.

3. Suppose we add noise to the data, $\tilde{Y} = \{y_i + \varepsilon_i\}_{i=1}^{n}$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. For which values of σ^2 will the result from part 1 hold with $> 95\%$ probability?

   Answer:
   This question was mistakenly added, please consider next question for a similar intuition.

4. Suppose we widen the boundaries of X to $(-2\pi, 2\pi)$. Write a short script to simulate samples $(X_i, \tilde{Y}_i)$ with different values of σ_i^2 and use 10-fold cross-validation to find corresponding optimal values $\hat{k}_i$. How do $\mathcal{L}(X_i, \hat{Y}_i, p)$ and $\hat{k}_i$ behave as k and σ^2 increase? Answer:

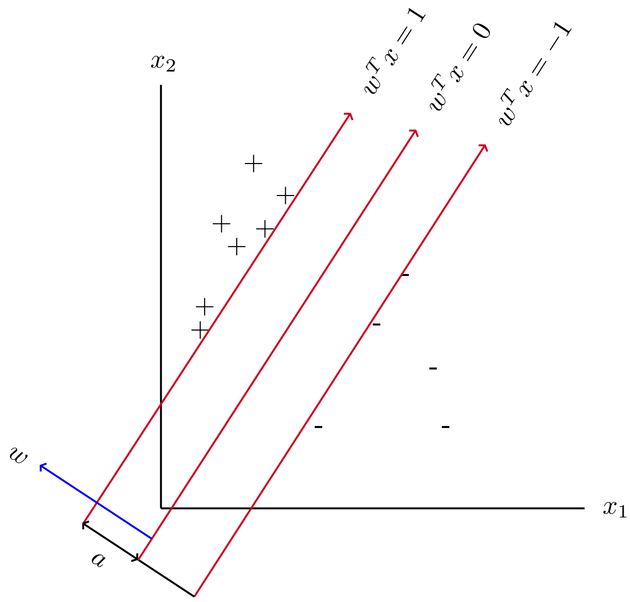
   As σ^2 begins to increase, we will and no longer have a special relation between our learned function and $sin(x)$. This means that we will no longer be guaranteed to get the more complex model when using cross-validation.



K=1
K=2
K=3
log10(test error)
-4
-6
-8
-10
-12
-14
0
1e-05
1e-04
1e-03
1e-2
Sigma

Problem 2 (Classification):

Consider the data set plotted below,



Show that $a = \frac{1}{||w||}$. How would L_2 regularization on w affect the margin around $w^T x = 0$?

Answer:
First we use the fact that the minimal euclidean distance from any point on a plane ($w^T x = b$) to the origin is $\frac{|b|}{||w||}$

We know that our other vectors are $w^T x + b = 1$ and $w^T x + b = -1$.

This is re-written as $w^T x = 1 - b$ and $w^T x = -1 - b$.

The distance between these two vectors is then $\frac{2}{||w||}$.

Therefore a is $\frac{1}{||w||}$.

We see that the margin around $w^T x = 0$ is maximized when $||w||$ is minimized.