Machine Learning - MT 2017
5 Basis Expansion, Learning Curves, Overfitting

Christoph Haase

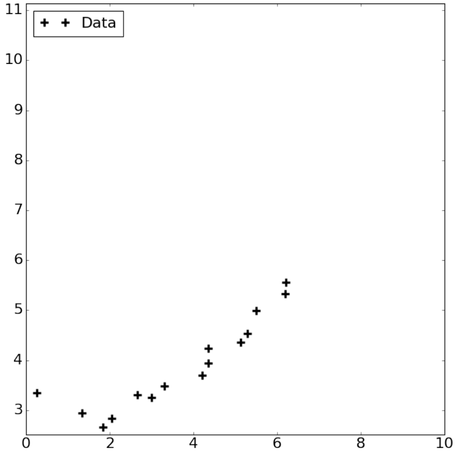
University of Oxford
October 18, 2017

Outline

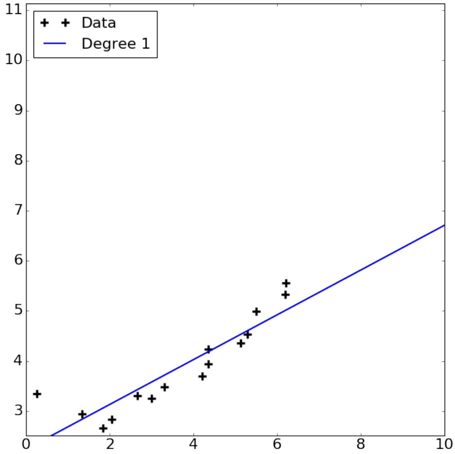
- Basis function expansion to capture non-linear relationships
- Understanding the bias-variance tradeoff
- How does overfitting occur

Outline

Linear Regression : Polynomial Basis Expansion



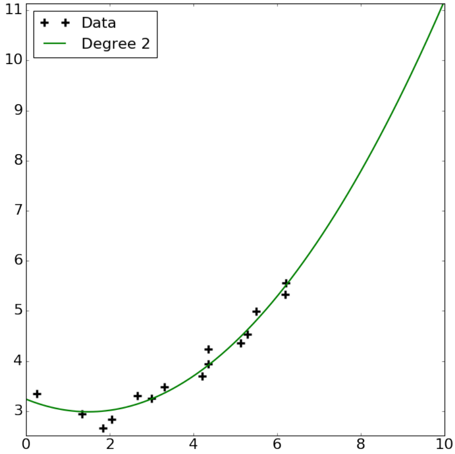
Linear Regression : Polynomial Basis Expansion



Linear Regression : Polynomial Basis Expansion

$$\boldsymbol{\phi}(x) = [1, x, x^2]$$

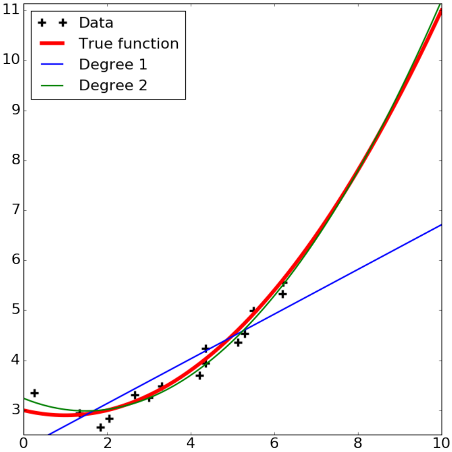
$$w_0 + w_1 x + w_2 x^2 = \boldsymbol{\phi}(x) \cdot [w_0, w_1, w_2]$$



Linear Regression : Polynomial Basis Expansion

$$\boldsymbol{\phi}(x) = [1, x, x^2]$$

$$w_0 + w_1 x + w_2 x^2 = \boldsymbol{\phi}(x) \cdot [w_0, w_1, w_2]$$

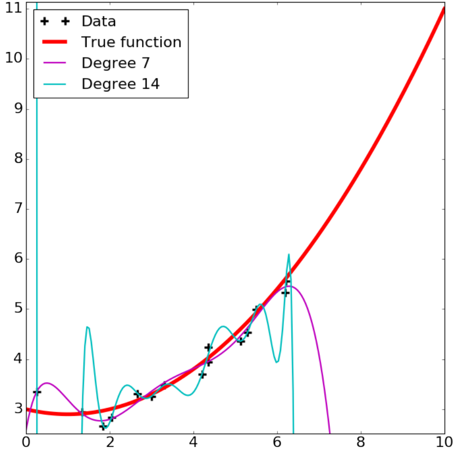


Linear Regression : Polynomial Basis Expansion

$\phi(x) = [1, x, x^2, \cdots, x^d]$

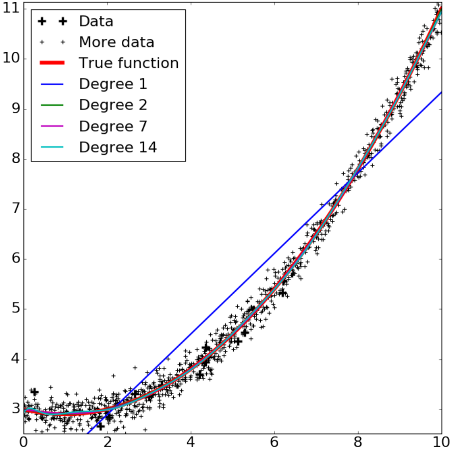
Model $y = \mathbf{w}^\mathsf{T}\phi(x) + \epsilon$

Here $\mathbf{w} \in \mathbb{R}^M$, where M is the number for expanded features



Linear Regression : Polynomial Basis Expansion

Getting more data can avoid overfitting!



Polynomial Basis Expansion in Higher Dimensions

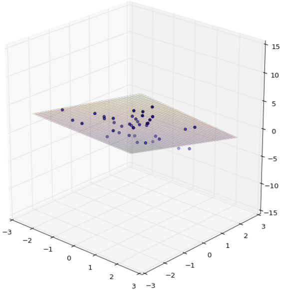
Basis expansion can be performed in higher dimensions

We're still fitting linear models, but using more features

$$y = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}) + \epsilon$$

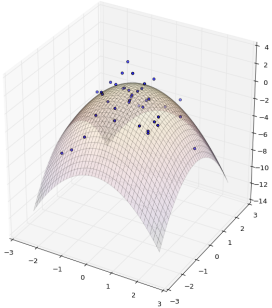
Linear Model

$\phi(\mathbf{x}) = [1, x_1, x_2]$



Quadratic Model

$\phi(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$



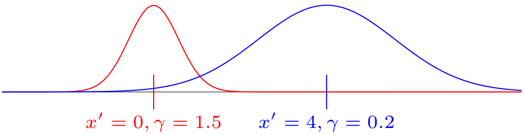
Using degree d polynomials in D dimensions results in $\approx D^d$ features!

Basis Expansion Using Kernels

We can use kernels as features

A Radial Basis Function (RBF) kernel with width parameter γ is defined as

$$\kappa(\mathbf{x}', \mathbf{x}) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$



A kernel computes the dot product $\kappa(\mathbf{x}', \mathbf{x}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R} = \phi(\mathbf{x}') \cdot \phi(\mathbf{x})$ of some expansion ϕ, see e.g. Sec. 5.7.2 in GBC.

Other kernels:

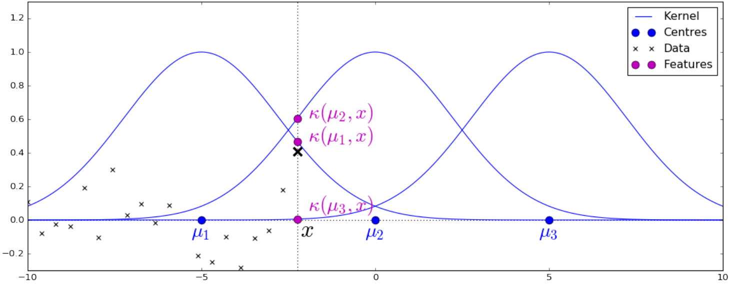
- Polynomial kernel: $\kappa(\mathbf{x}', \mathbf{x}) = (\mathbf{x}^{\mathsf{T}} \cdot \mathbf{x}' + c)^d$
- String kernels
- Graph kernels

Basis Expansion Using Kernels

- RBF kernel: $\kappa(\mathbf{x}', \mathbf{x}) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$
- Choose centres $\mu_1, \mu_2, \ldots, \mu_M$
- Feature map: $\phi(\mathbf{x}) = [1, \kappa(\mu_1, \mathbf{x}), \ldots, \kappa(\mu_M, \mathbf{x})]$

$$y = w_0 + w_1 \kappa(\mu_1, \mathbf{x}) + \cdots + w_M \kappa(\mu_M, \mathbf{x}) + \epsilon = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}) + \epsilon$$

- How do we choose the centres?



Basis Expansion Using Kernels

One reasonable choice is to choose data points themselves as centres for kernels

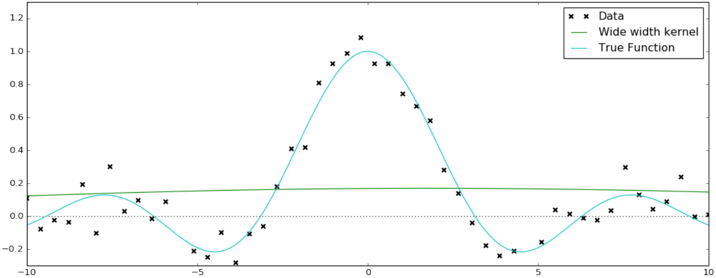
Need to choose width parameter γ for the RBF kernel

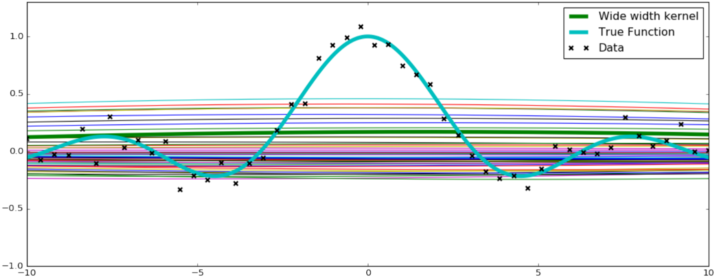
$$\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$

As with the choice of degree in polynomial basis expansion depending on the width of the kernel overfitting or underfitting may occur

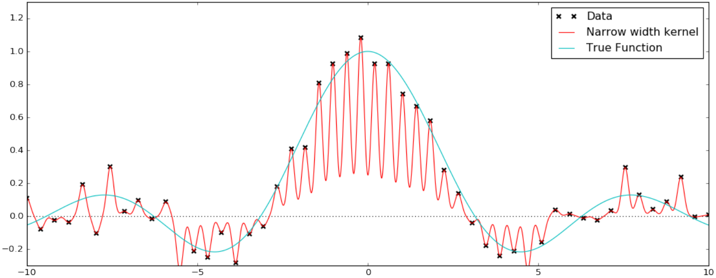
- Overfitting occurs if the width is too small, *i.e.,* γ very large
- Underfitting occurs if the width is too large, *i.e.,* γ very small

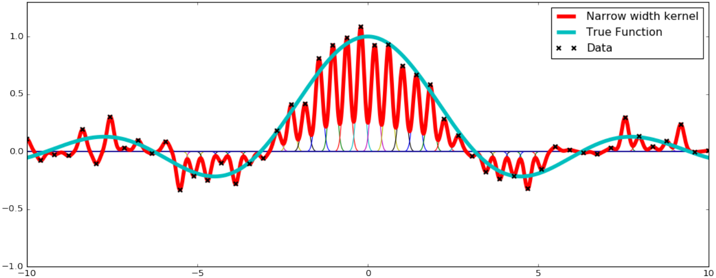
When the kernel width is too large



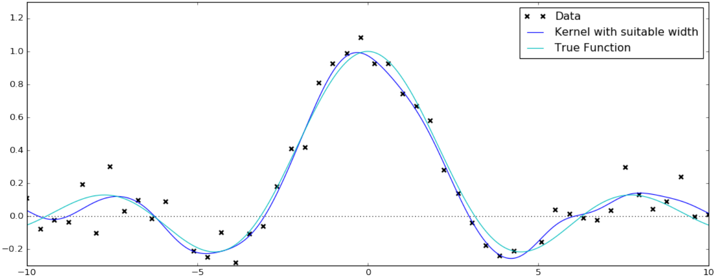


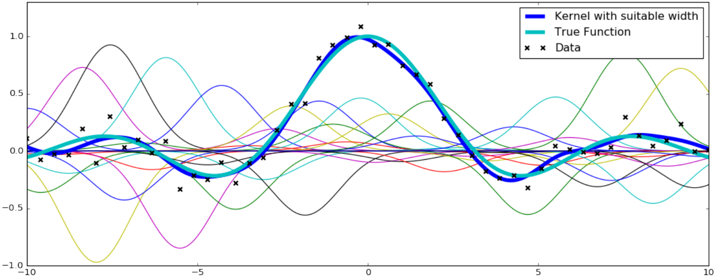
When the kernel width is too small



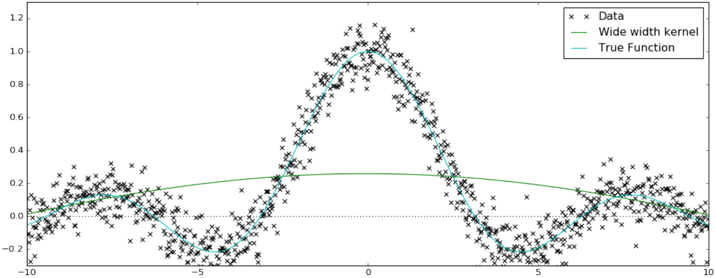


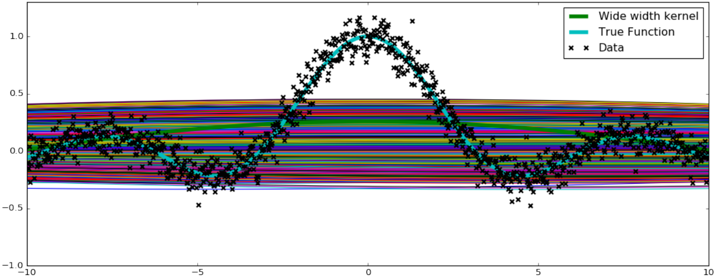
When the kernel width is chosen suitably



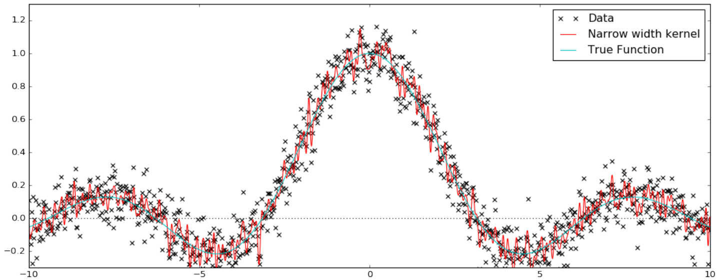


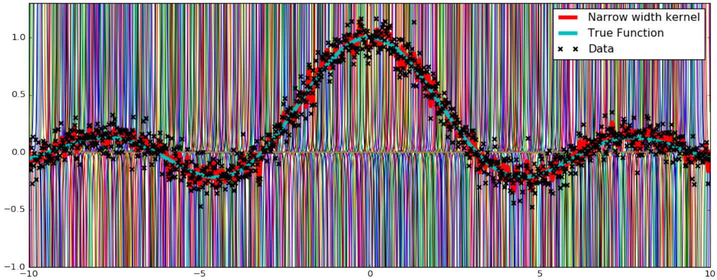
Big Data: When the kernel width is too large



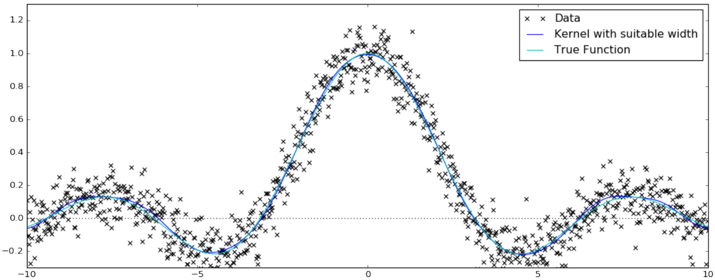


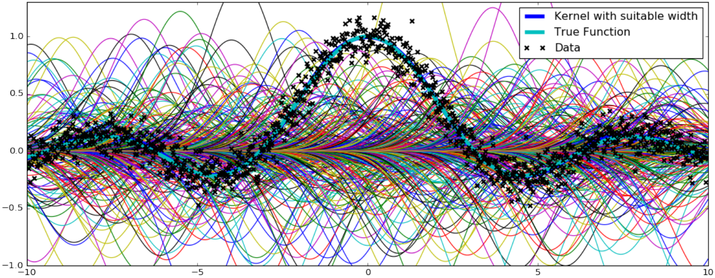
Big Data: When the kernel width is too small





Big Data: When the kernel width is chosen suitably





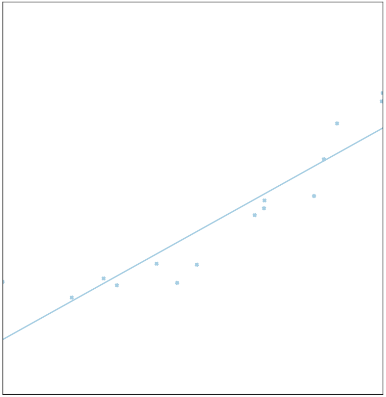
Basis Expansion using Kernels

- Overfitting occurs if the kernel width is too small, *i.e.*, γ very large
  - Having more data can help reduce overfitting!
- Underfitting occurs if the width is too large, *i.e.*, γ very small
  - Extra data does not help at all in this case!
- When the data lies in a high-dimensional space we may encounter the curse of dimensionality
  - If the width is too large then we may underfit
  - Might need exponentially large (in the dimension) sample for using modest width kernels
  - Connection to Problem 1 on Sheet 1

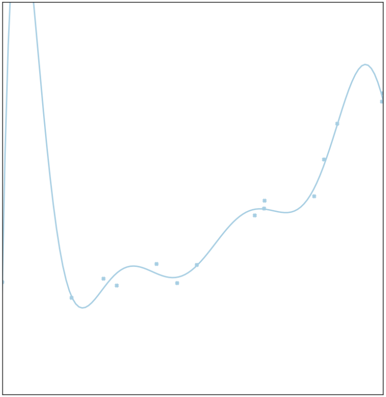
Outline

The Bias Variance Tradeoff

High Bias

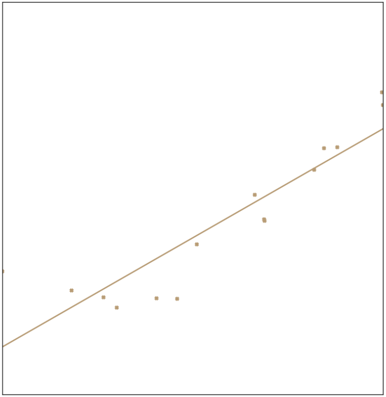


High Variance

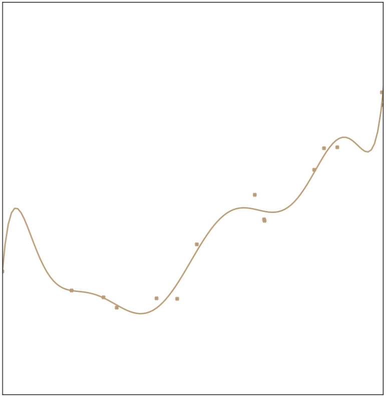


The Bias Variance Tradeoff

High Bias

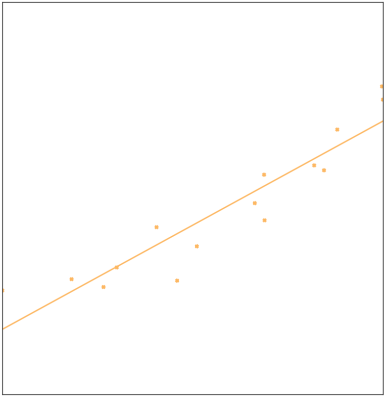


High Variance

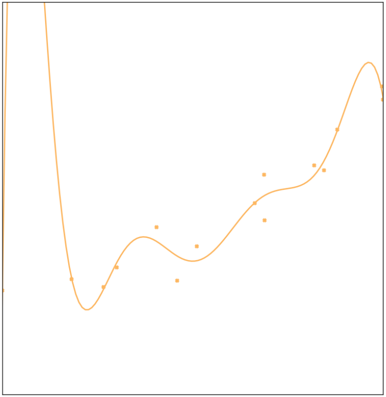


The Bias Variance Tradeoff

High Bias

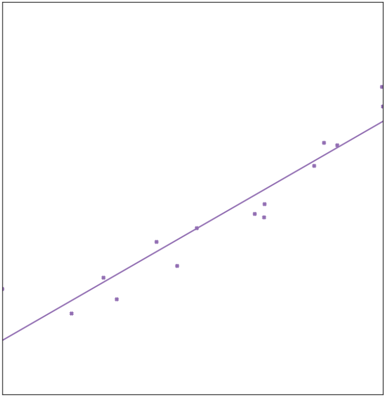


High Variance

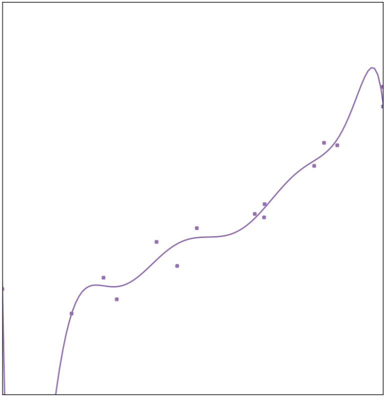


The Bias Variance Tradeoff

High Bias

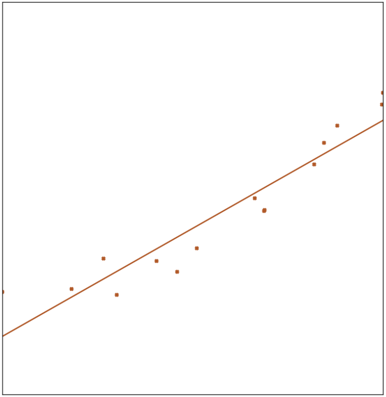


High Variance

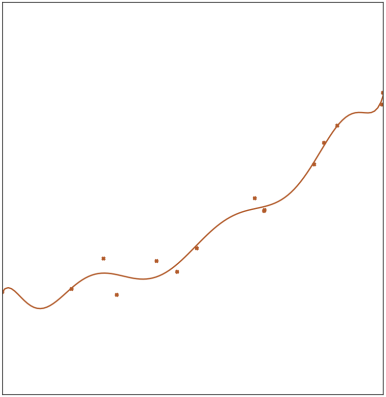


The Bias Variance Tradeoff

High Bias

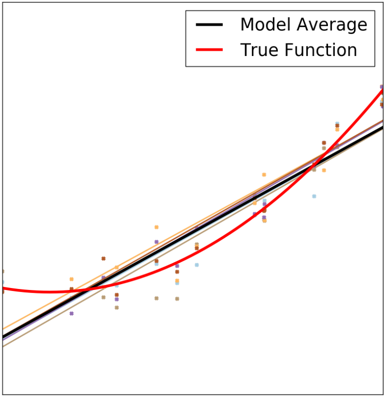


High Variance

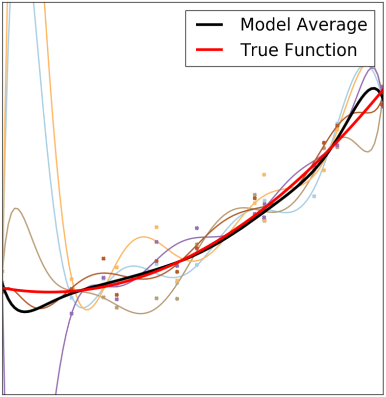


The Bias Variance Tradeoff

High Bias



High Variance



The Bias Variance Tradeoff

- Having high bias means that we are underfitting
- Having high variance means that we are overfitting
- The terms bias and variance in this context are precisely defined statistical notions
- See Problem Sheet 2, Q3 for precise calculations in one particular context
- See Sec. 5.4 in the GBC book for a much more detailed description

Learning Curves

Suppose we've trained a model and used it to make predictions

But in reality, the predictions are often poor

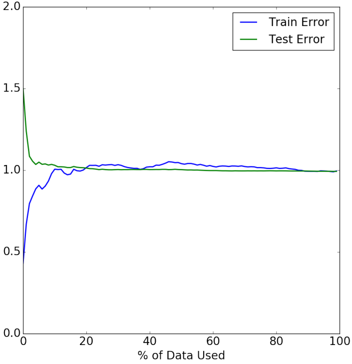
- How can we know whether we have high bias (underfitting) or high variance (overfitting) or neither?
  - Should we add more features (higher degree polynomials, lower width kernels, etc.) to make the model more expressive?
  - Should we simplify the model (lower degree polynomials, larger width kernels, etc.) to reduce the number of parameters?
- Should we try and obtain more data?
  - Often there is a computational and monetary cost to using more data

Learning Curves

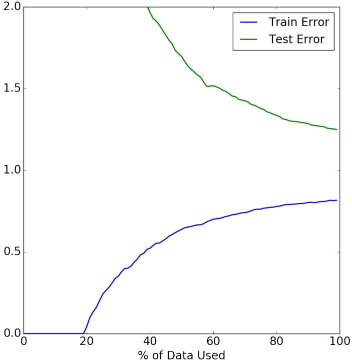
Split the data into a training set and testing set

Train on increasing sizes of data

Plot the training error and test error as a function of training data size



More data is not useful



More data would be useful

Outline

Overfitting: How does it occur?

When dealing with high-dimensional data (which may be caused by basis expansion) even for a linear model we have many parameters

With $D = 100$ input variables and using degree 10 polynomial basis expansion we have $\sim 10^{20}$ parameters!

Enrico Fermi to Freeman Dyson

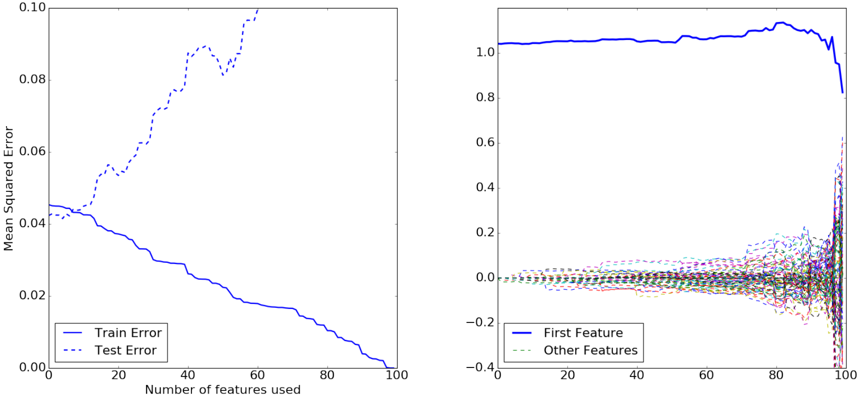
"I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." [video]

Overfitting: How does it occur?

Suppose we have $D = 100$ and $N = 100$ so that $\mathbf{X}$ is 100×100

Suppose every entry of $\mathbf{X}$ is drawn from $\mathcal{N}(0, 1)$

And let $y_i = x_{i,1} + \mathcal{N}(0, \sigma^2)$, for $\sigma = 0.2$



Next Time (Lecture Theatre A)

Coping with overfitting:

- Ridge Regression and Lasso
- Model Selection for tuning hyperparameters