Lecture 2: DL Basics

Deep Learning (深度学习)

Overview

- Linear Algebra
- Probability and Information Theory
- Mathematical Optimization
- Machine Learning Basics

- All these materials can refer to the references books

Overview

- Linear Algebra
- Probability and Information Theory
- Mathematical Optimization
- Machine Learning Basics

- All these materials can refer to the references books

Machine Learning Basics

Overview

- Introduction to ML
- Capacity, Overfitting and Underfitting
- Estimators, Bias and Variance
  - Maximum Likelihood Estimation
  - Bayesian Statistics
- Challenges Motivating Deep Learning

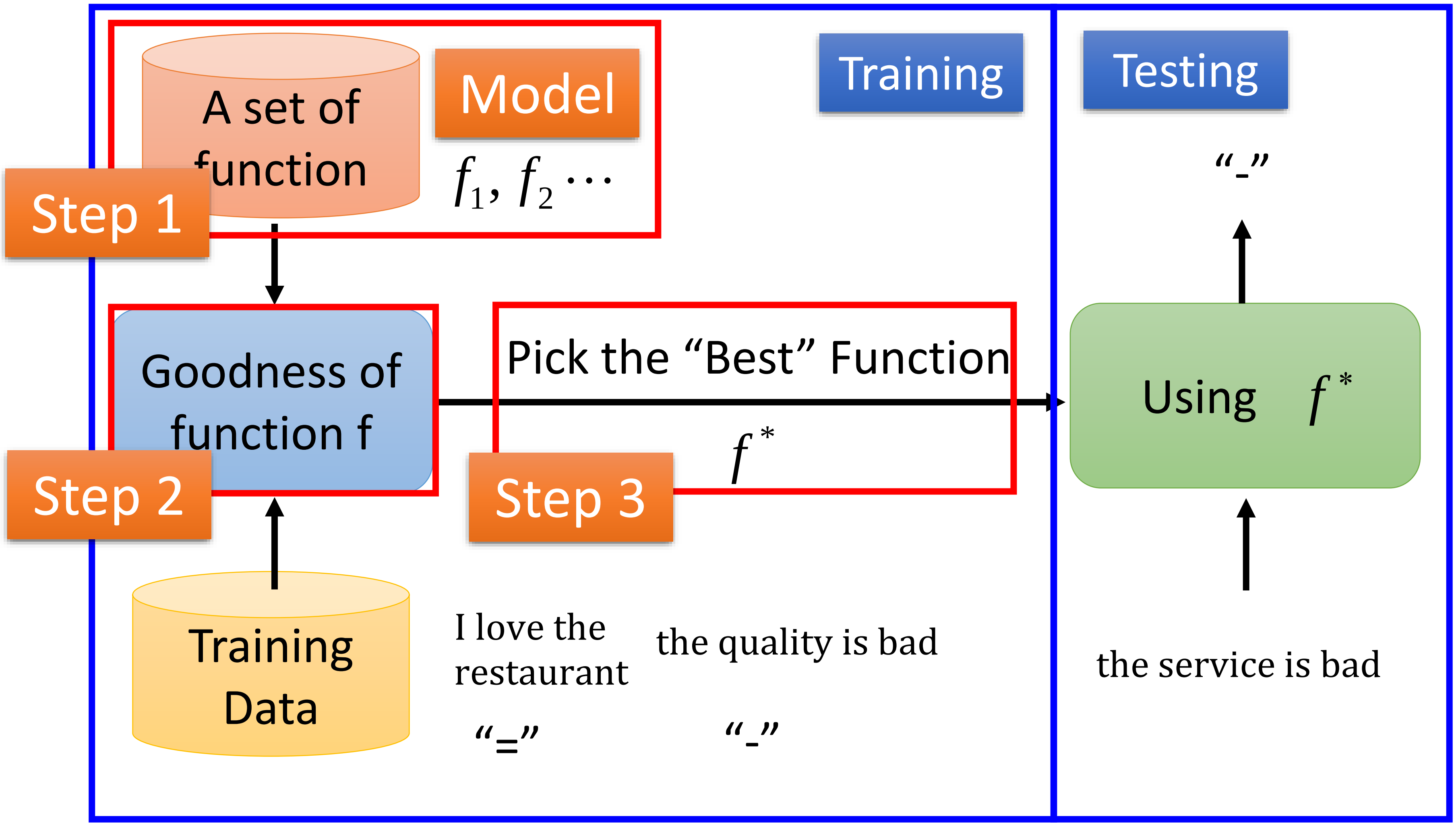
Learning Algorithms

- An algorithm that is able to learn from data
- Mitchell (1997)
  - "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measures *P*, if its performance at tasks in *T*, as measured by *P*, improved with experience *E*."

Framework

Sentiment analysis:

f("I love the restaurant") = "+" (positive)



Three Steps for Machine Learning

Step 1: define a set of function → Step 2: goodness of function → Step 3: pick the best function

model

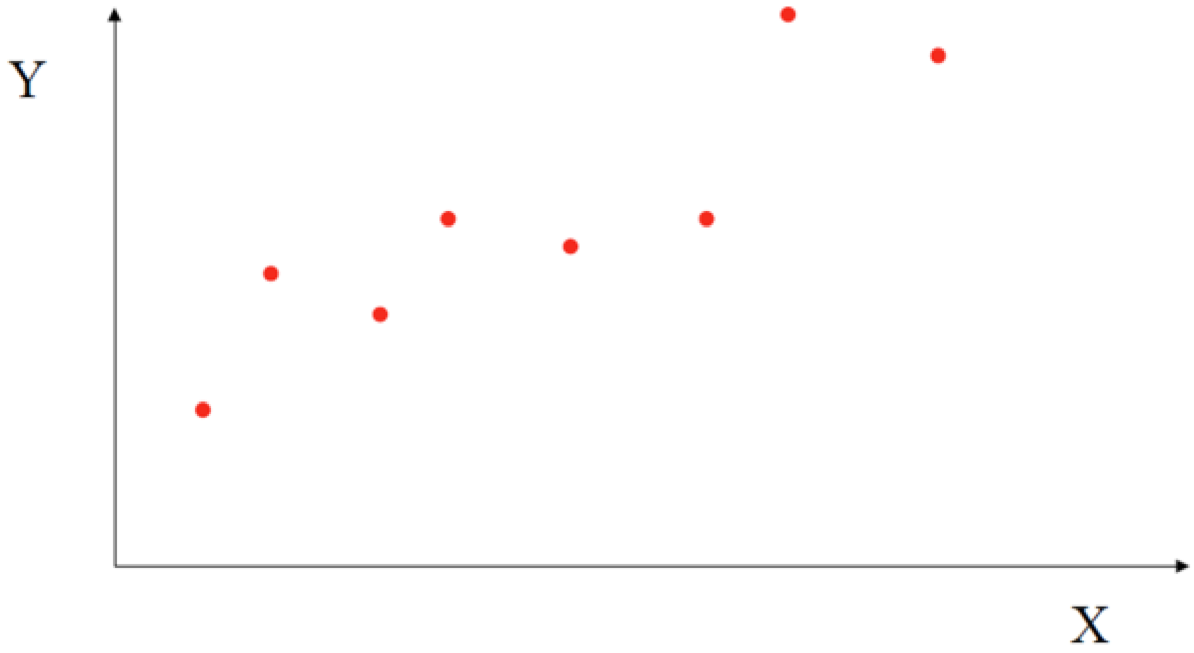
loss function

optimization

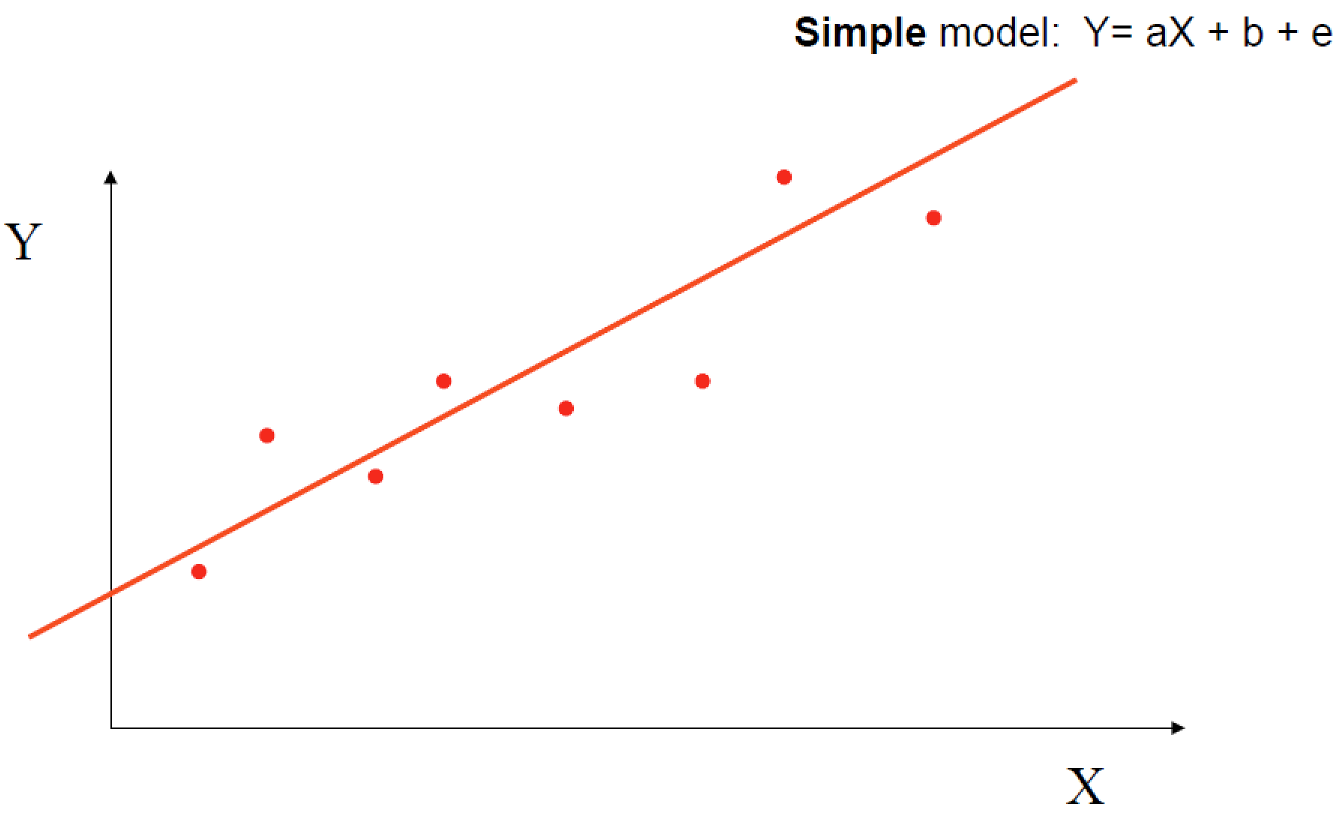
Capacity, Overfitting and Underfitting

- Generalization
  - The ability to perform well on previously unobserved inputs (i.e. out-of-sample)
- Data generating process
  - $i.i.d.$ assumptions = independently and identically distributed
  - Data-generating distribution, p_{data}
  - Expected [Generalization error (or test error)] = Expected (training error)
- Goal of ML algorithms
  - Make the training error small
    - If not, underfitting
  - Make the gap between training and test error small
    - If not, overfitting

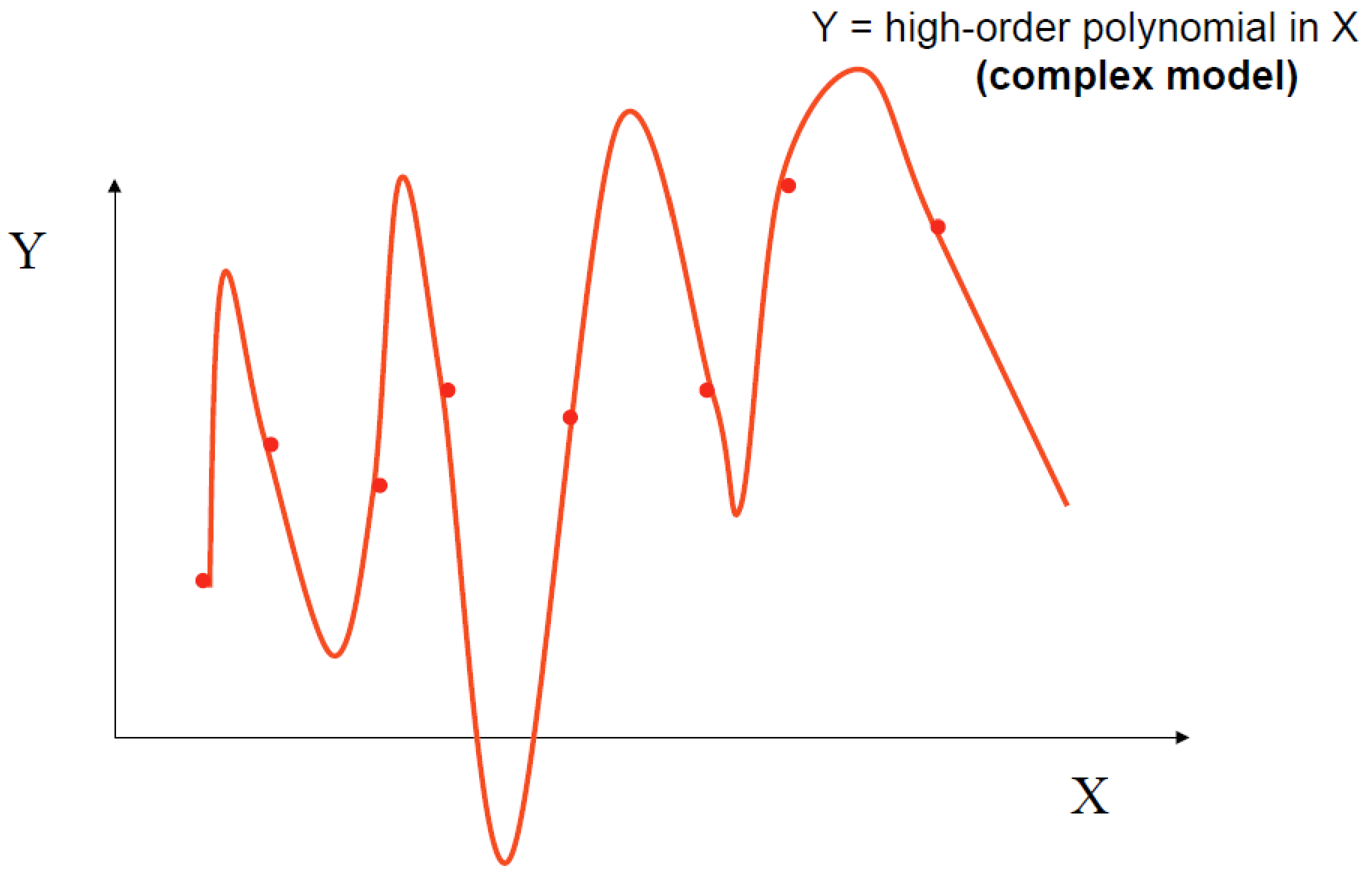
Overfitting and Complexity



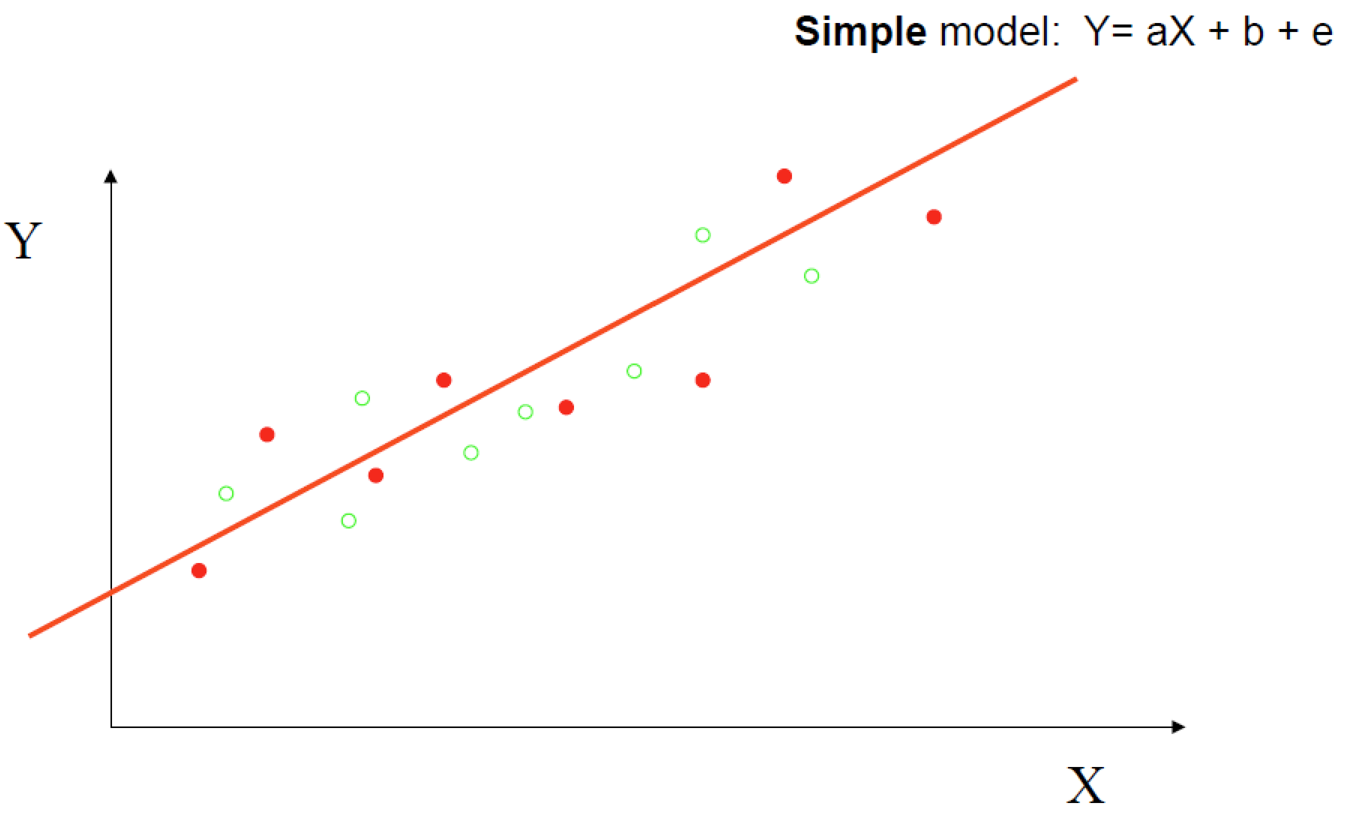
Overfitting and Complexity



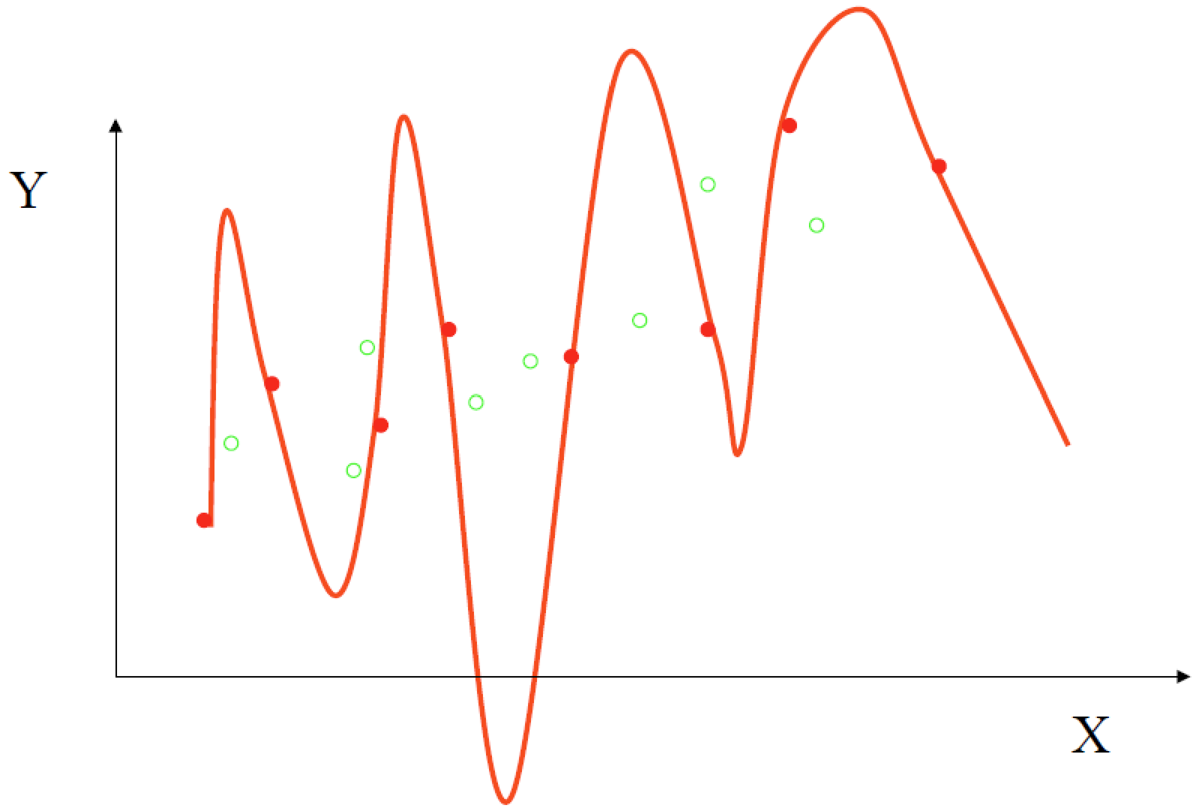
Overfitting and Complexity



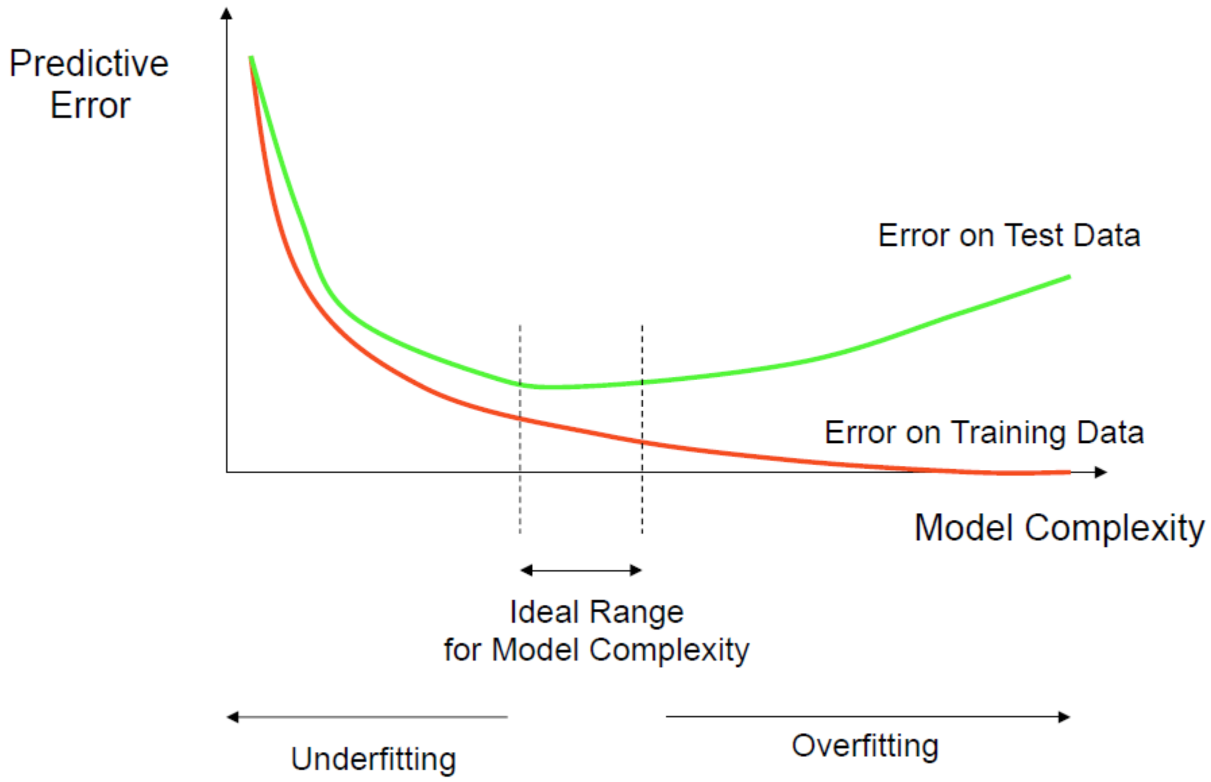
Overfitting and Complexity



Overfitting and Complexity

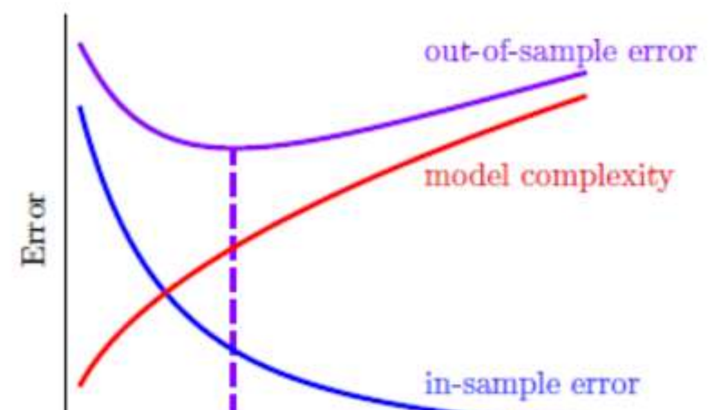


How Overfitting affects Prediction

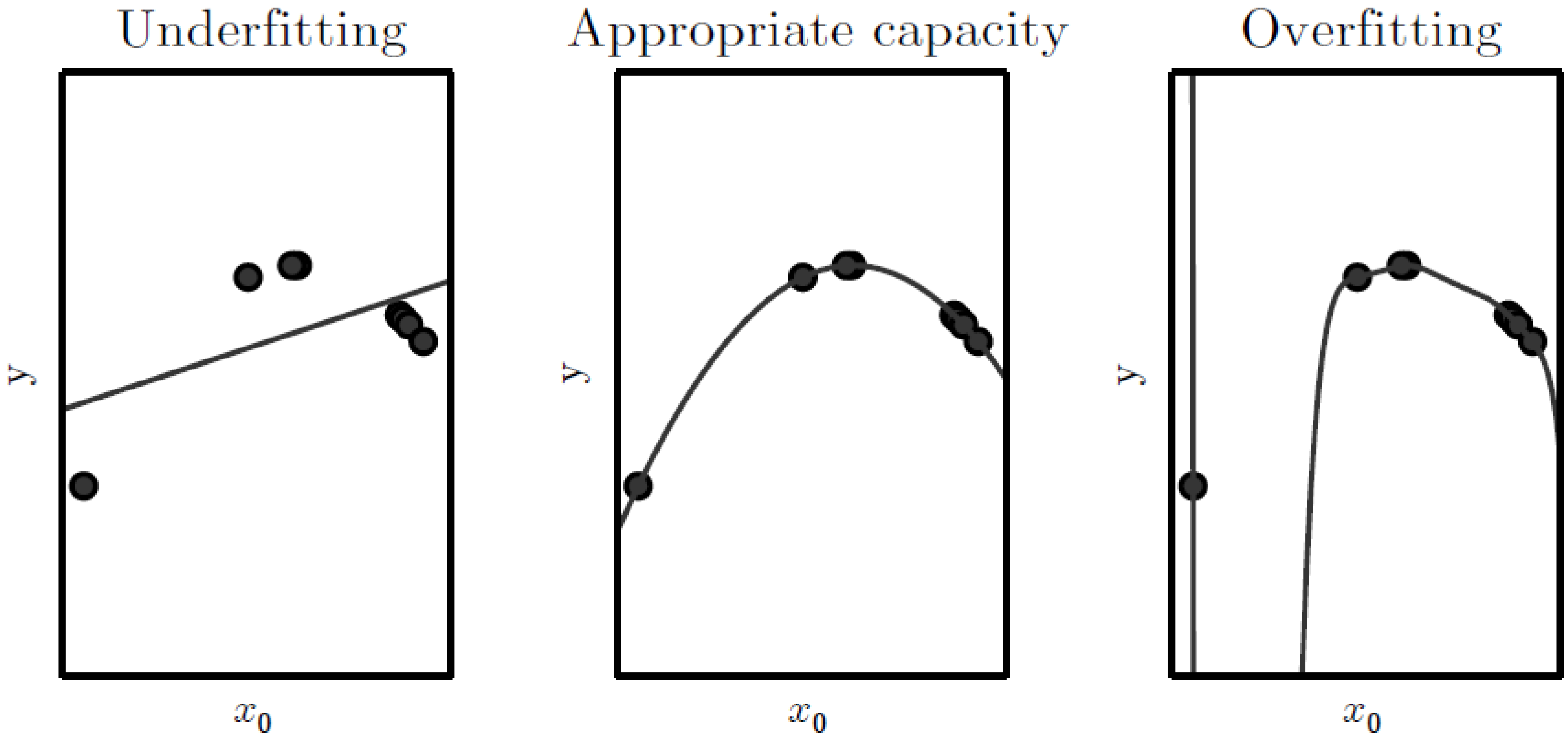


Capacity

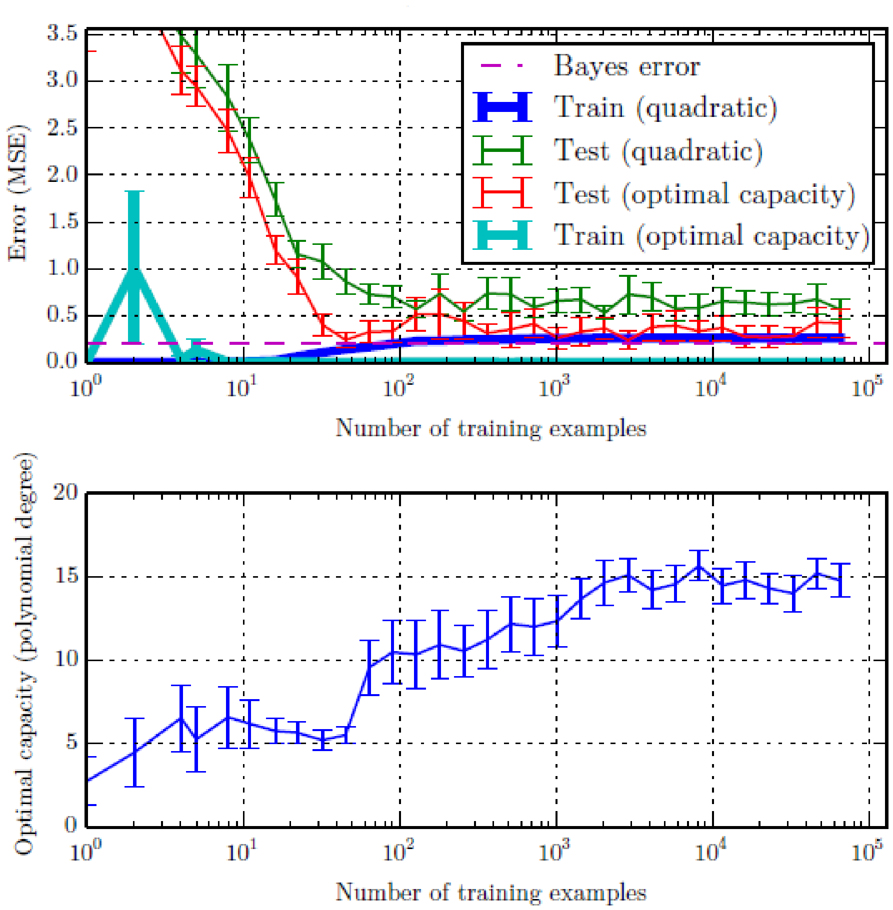
- A model’s ability to fit a wide variety of functions
- Ways to control the capacity
  - Hypothesis space (input features)
  - The model
    - Representation capacity vs. effective capacity
    - Occam’s razor
      - Quantifying model capacity (VC dimension)
    - Nonparametric vs. parametric
  - Size of the training set



Polynomial Estimation



Training Data Size



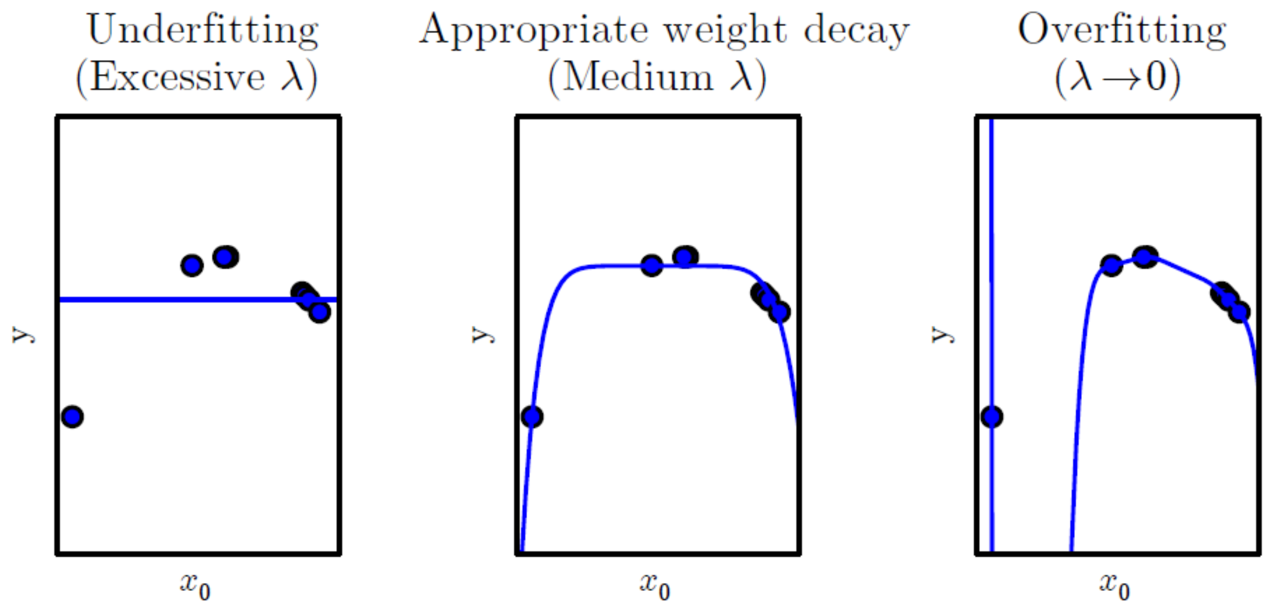
Regularization

- Cost function

$$J(w) = MSE_{train}$$

- Cost function + penalty (regularizer)

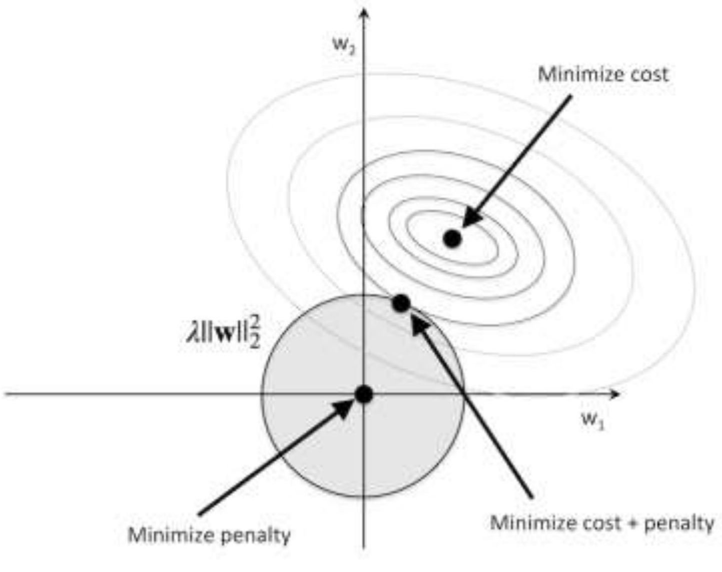
$$J(w) = MSE_{train} + \lambda f(w)$$

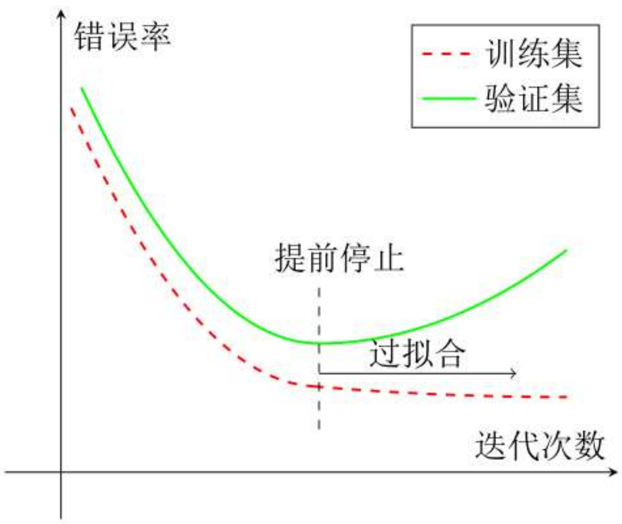


Regularization

所有损害优化的方法都是正则化。

- 增加优化约束
  - L1/L2约束、数据增强
- 干扰优化过程
  - 权重衰减、随机梯度下降、提前停止





No free Lunch Theorem

- No machine learning algorithm is universally better than any other
  - The most sophisticated algorithm has the same average performance (over all possible tasks) as merely predicting that every point belongs to the same class
  - Goal of real ML research is to understand the mapping of ML algorithms to data generating distributions

Estimators, Bias and Variance

Point Estimation

- Any function of the data, $\{x^1, \ldots, x^m\}$ a set of m i.i.d. data points

$$\hat{\theta}_m = g(x^1, \ldots, x^m)$$

- Function estimation
  - Point estimator in function space, e.g.
  - $y = f(x) + \epsilon$

Bias

- $\text{bias}(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta$
- Unbiased: $\text{bias}(\hat{\theta}_m) = 0$
- Asymptotically unbiased: $\lim_{m \to \infty} \text{bias}(\hat{\theta}_m) = 0$
- Examples
  - Bernoulli distribution
  - Gaussian Distribution Estimators of the mean and variance

Variance and Standard Error

- Variance of an estimator

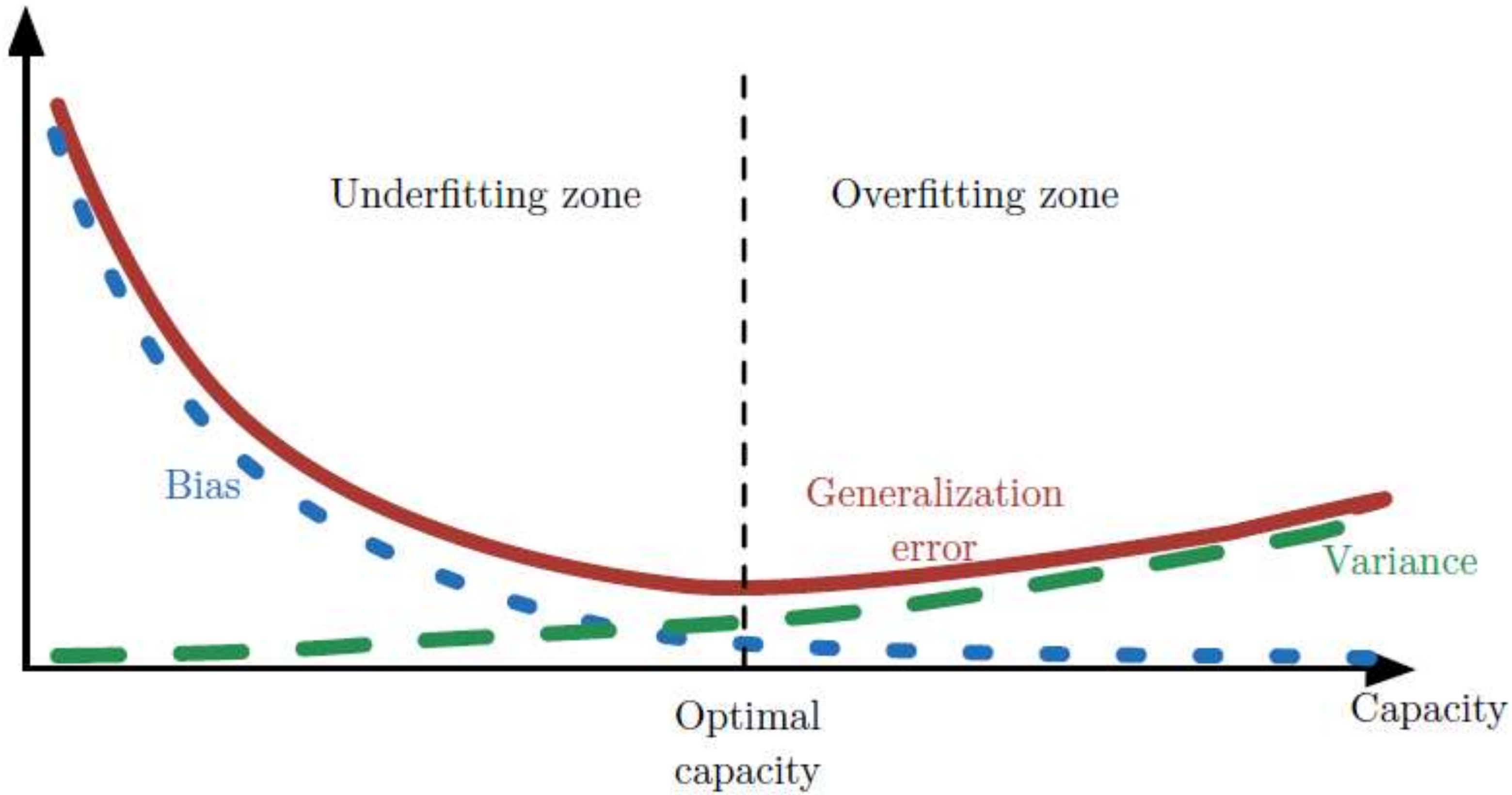
$$\mathrm{var}(\hat{\theta})$$

  - Variance of the estimator as we independently resample the dataset from the underlying data-generating process
  - Standard error: $\mathrm{SE}(\hat{\theta})$
  - Central limit theorem: normal distribution
    - 95% confidence interval centered on the mean $\hat{\mu}_m$

$$(\hat{\mu}_m - 1.96\mathrm{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\mathrm{SE}(\hat{\mu}_m))$$

Tradeoff Between Bias and Variance

$$MSE = \mathbb{E}\left[\left(\hat{\theta}_m - \theta\right)^2\right] = Bias\left(\hat{\theta}_m\right)^2 + Var\left(\hat{\theta}_m\right)$$



Consistency

- $\text{plim}_{m\to\infty} \hat{\theta}_m = \theta$
- $\forall \epsilon > 0, P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$, as $m \to \infty$
- The bias diminishes as the increase of data size
    - The reverse is not true

MLE

$$\theta_{ML} = \arg\max_{\theta} p_{model}(\mathbb{X};\theta)$$
$$= \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(x^i;\theta)$$

- Take the logarithm

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{model}(x^i;\theta)$$
$$= \arg\max_{\theta} \mathbb{E}_{x\sim\hat{p}_{data}} \log p_{model}(x;\theta)$$

KL Explanation

$$D_{KL}(\hat{p}_{data} \parallel p_{model}) = \mathbb{E}_{x \sim \hat{p}_{data}}[\log \hat{p}_{data}(x) - \log p_{model}(x)]$$

- To minimize the KL divergence, equal to minimize

$$-\mathbb{E}_{x \sim \hat{p}_{data}}[\log p_{model}(x)]$$

Conditional Log-likelihood

- $\theta_{\text{ML}} = \arg\max_{\theta} \prod_{i=1}^{m} \log P\left(y^{i} \mid x^{i}; \theta\right)$
- Example
  - Linear regression as Maximum Likelihood

Properties of ML

- The best estimator asymptotically in terms of convergences as m increases
  - Consistency
  - Efficiency
- Property of consistency
  - p_{data} must lie within the model family $p_{model}(.\,;\theta)$
  - p_{data} must correspond to exactly one value of θ

Bayesian Statistics

- Consider all possible value of θ when making a prediction
- $p(\theta|x^1,\ldots,x^m) = \frac{p\left(x^1,\ldots,x^m\middle|\theta\right)p(\theta)}{p(x^1,\ldots,x^m)}$
  - Prior probability distribution: $p(\theta)$ (high entropy to reflect high uncertainty)
  - Data likelihood: $p(x^1,\ldots,x^m|\theta)$
- Major differences with MLE
  - Make prediction using full distribution over θ

$$p(x^{m+1}|x^1,\ldots,x^m) = \int p(x^{m+1}|\theta)\,\mathrm{p}(\theta|x^1,\ldots,x^m)d\theta$$

  - The influence of priors
- Example: Bayesian Linear Regression

Maximum A Posteriori Estimation (MAP)

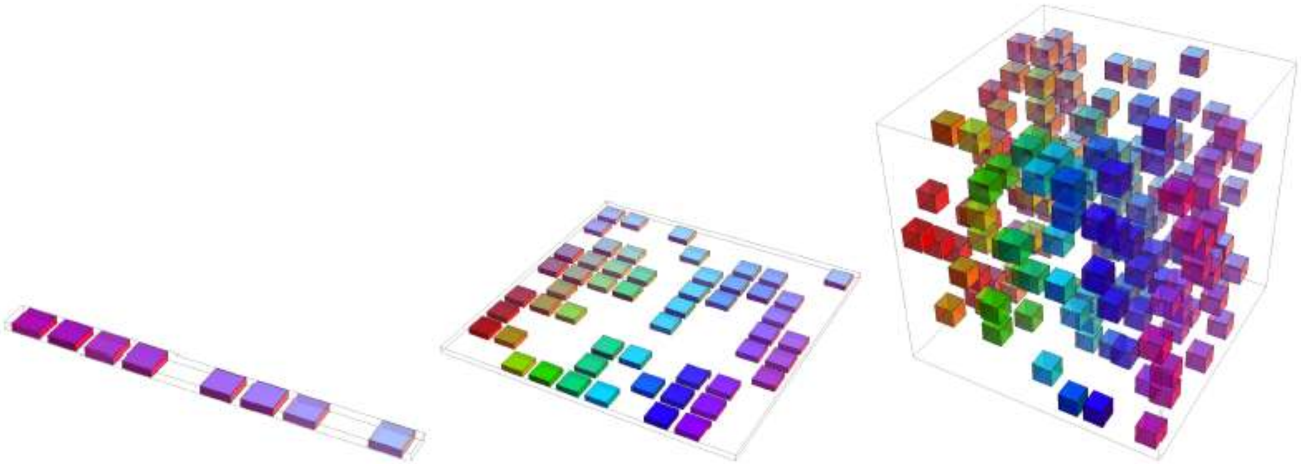
$$\theta_{MAP} = \arg\max_{m} p(\theta|x)$$
$$= \arg\max_{m} \log p(\theta|x) + \log p(\theta)$$

- Advantages:
  - With full Bayesian, leverage information brought by prior and cannot be found in training data, reduce variance but increase bias
  - Could design complicated yet interpretable regularization terms
    - MLE + regularizer = MAP

Challenges Motivating Deep Learning

The Curse of Dimensionality

- ML learning becomes exceedingly difficult when the number of dimensions in the data is high
  - Statistical challenge



  - Arose the smoothness assumption

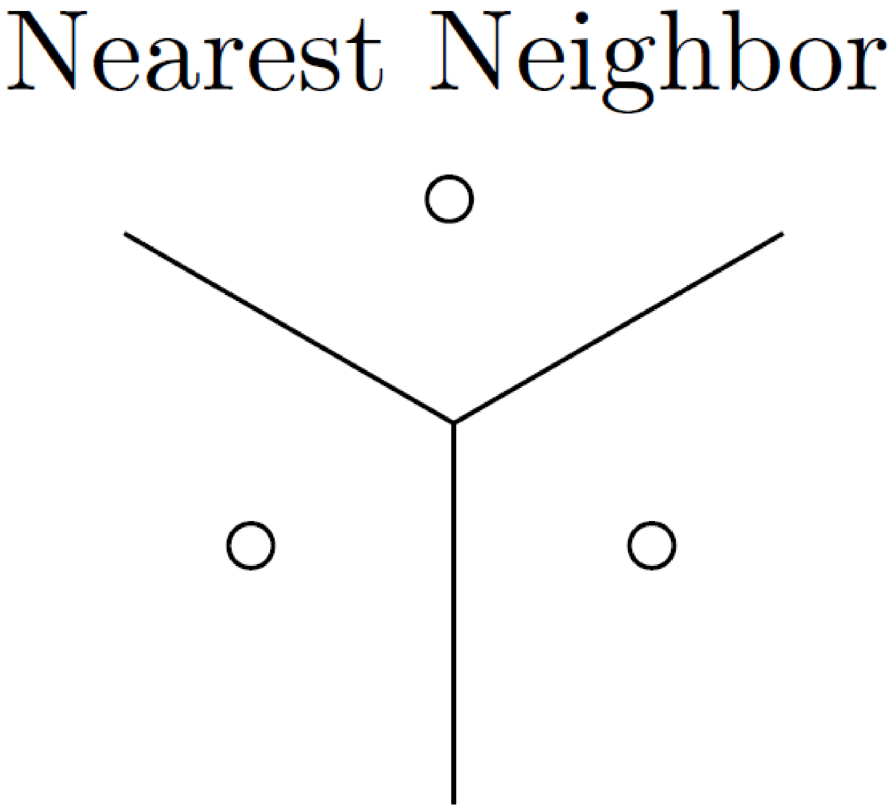
Local Constancy and Smoothness Regularization

- Local constancy prior: Learnt function should keep stable within a small region

$$f^*(x) \approx f^*(x + \epsilon)$$

- Many simpler algorithms rely exclusively on the local constancy prior to generalize well
  - fail to scale to the statistical challenges in AI-level tasks
    - E.g. KNN, decision tree

Break Input Space Into Regions

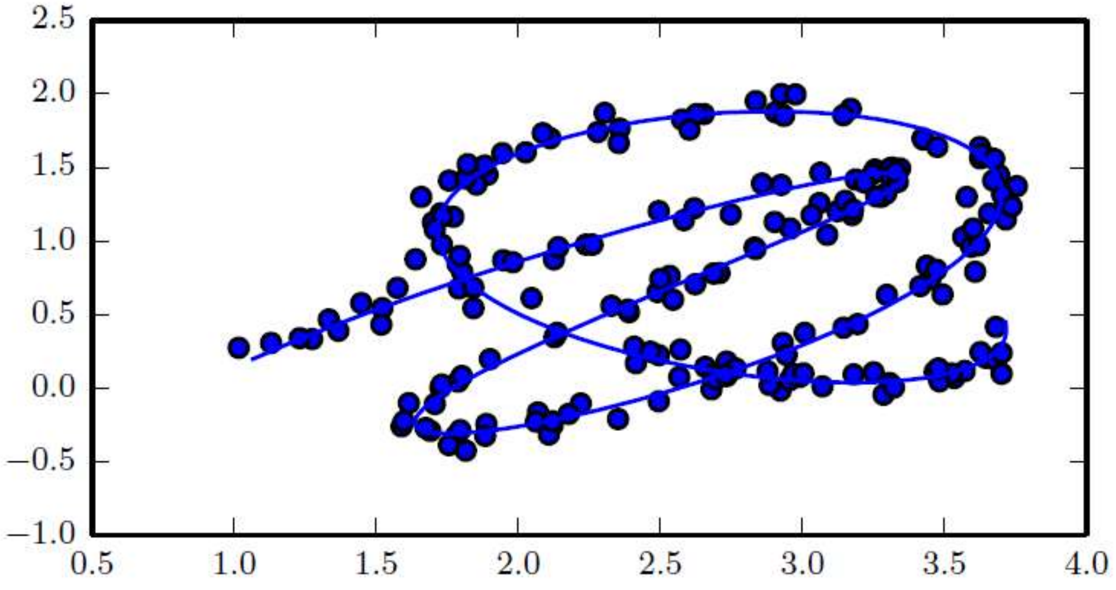


Local Constancy and Smoothness Regularization

- To answer two questions
  - Whether possible to represent a complicated function efficiently?
  - Whether possible to generalize well to new inputs?
- Solutions
  - Introduce dependencies among regions
    - DL methods DO without stronger task specific assumptions: exponential gain

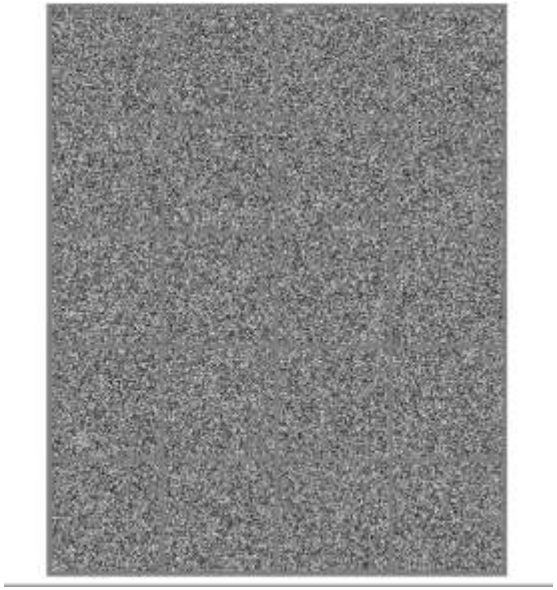
Manifold Learning

- Manifold assumption
  - Most of $\mathbb{R}^n$ consists of invalid inputs
  - Interesting variations happen only when move from one manifold to another
  - The data lies along a low-dimensional manifold



Manifold Learning

- Images, sounds and text strings are highly concentrated, and in favor of manifold hypothesis
  - Represent data in terms of coordinates on the manifold
- Manifold transformations are imaginably possible



Manifold Learning

- Extracting manifolds is challenging but promising
  - E.g. textbook section 20.10.4

Reading Materials

- Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer Publisher, 2006