Basic learning principles

229351

Main principle

- Predictors $X = (X_1, X_2, \ldots, X_p)$
- Response Y

Main principle

- Predictors $X = (X_1, X_2, \ldots, X_p)$
- Response Y

Assumption: There's some function f such that

$$Y = f(X)$$

Goal: Find an estimate $\hat{f}$ of f.

Frequentism

- We would like to say that $\hat{f}$ is accurate "on average".
    - or with "high probability".

- Where does the **randomness** come from?

Frequentism

- We would like to say that $\hat{f}$ is accurate “on average”.
    - or with “high probability”.

- Where does the **randomness** come from?

- The frequentist idea: we don’t have one fixed dataset, but several datasets sampled from a population.
    - The randomness comes from the sampling process.

Hypothesis testing

- We ask a question of whether the data fall under a specified distribution. This is the null hypothesis
  - and also the alternative hypothesis.

Hypothesis testing

- We ask a question of whether the data fall under a specified distribution. This is the null hypothesis
  - and also the alternative hypothesis.

- We then collect data and create a rule that measures how well the data fit the null distribution.
  - If not so much, then the rule rejects the null.

Hypothesis testing

- We ask a question of whether the data fall under a specified distribution. This is the null hypothesis
  - and also the alternative hypothesis.
- We then collect data and create a rule that measures how well the data fit the null distribution.
  - If not so much, then the rule rejects the null.
- In theory, the rule must tell us the correct answer with high probability.
  - For example, 5% or 1% of rejecting the null when data came from the null distribution.

Making decisions

- Data X is sampled from a probability distribution, indexed by parameter θ.

Making decisions

- Data X is sampled from a probability distribution, indexed by parameter θ.

- Compute an estimator $\delta(X)$ that we will use to estimate θ.

Making decisions

- Data X is sampled from a probability distribution, indexed by parameter θ.

- Compute an estimator $\delta(X)$ that we will use to estimate θ.

- Measure the performance of $\delta(X)$ using a loss function:

$$\ell(\theta, \delta(X))$$

  - Ex: $\ell(\theta, \delta(X)) = 0$ if $\delta(X) = \theta$, 1 otherwise.

Risk function

- There is randomness in the loss function

$$\ell(\theta, \delta(X)).$$

- We would like to know the performance of $\delta(X)$ on average, hence the risk function:

$$R(\theta, \delta) = \mathbb{E}[\ell(\theta, \delta(X)]$$

Risk under squared loss

- The squared loss: $\ell(\theta, \delta(X)) = (\delta(X) - \theta)^2$.
- Expanding out the risk:

$$R(\theta, \delta) = \mathbb{E}[\ell(\theta, \delta(X))]$$

The cross-product

$$2\mathbb{E}[\delta(X) - \mathbb{E}[\delta(X)])(\mathbb{E}[\delta(X)] - \theta)]$$

Risk under squared loss

$$\begin{aligned} R(\theta, \delta) = \mathbb{E}[(\delta(X) - \mathbb{E}[\delta(X)])^2] \\ + 2\mathbb{E}[\delta(X) - \mathbb{E}[\delta(X)])(\mathbb{E}[\delta(X)] - \theta)] \\ + \mathbb{E}[(\mathbb{E}[\delta(X)] - \theta)^2] \end{aligned}$$

Bias-variance trade-off

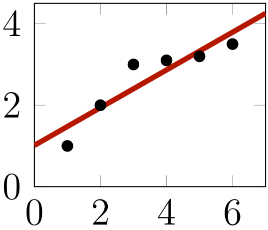
$$R(\theta, \delta) = \text{Variance} + \text{Bias}^2$$

- When one adjusts the model δ to decrease the variance, the bias increases and vice versa.

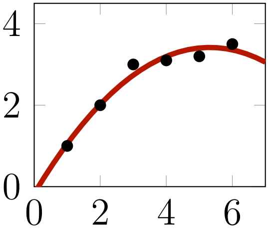
- We can turn this into a constrained optimization problem
  - Example: find an estimator that minimizes the variance, while being unbiased.

Polynomial regressions

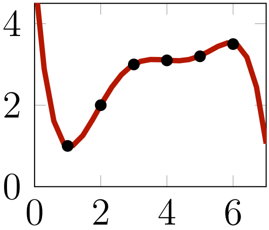
Polynomial regressions under different degrees



d=1



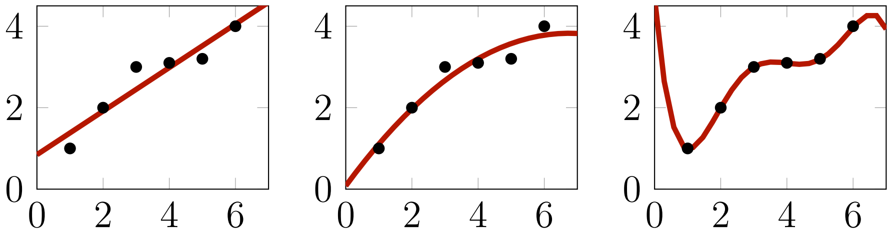
d=2



d=5

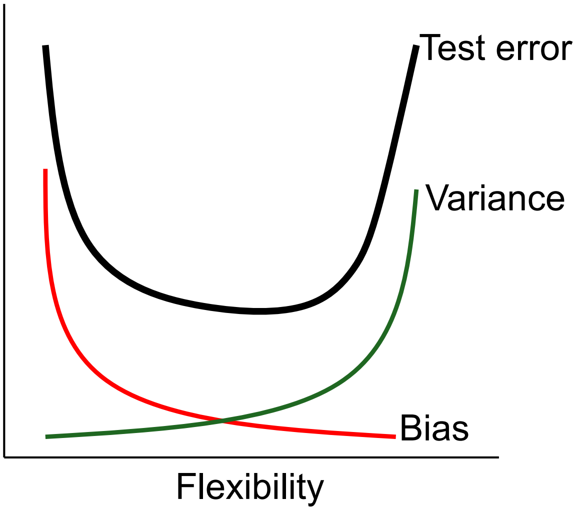
Polynomial regressions

The last point is moved up by 0.5.



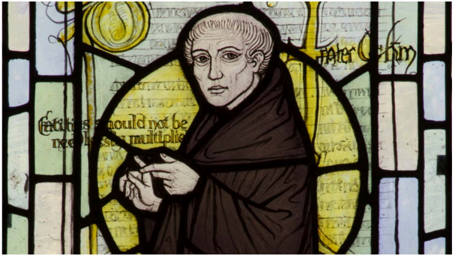
d=1 d=2 d=5

Bias-Variance tradeoff



underfitting vs overfitting

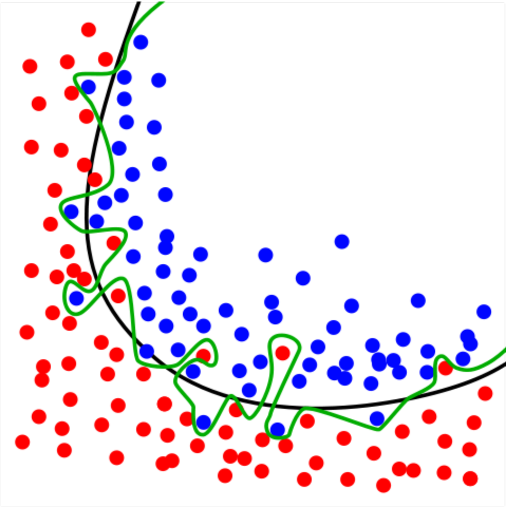
Occam's razor



- William of Ockham (1287-1347), a theologian and philosopher
- "The simplest explanation is usually the right one."

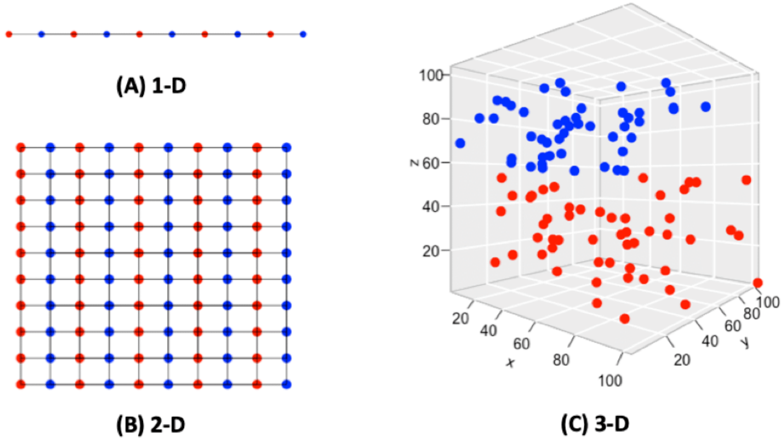
No free lunch's theorem

There's no model that works well on every problems.



Curse of dimensionality

Number of data points to learn the distribution grows exponentially with the dimension.



Supervised learning

Learning problem:

$$Y = f(X)$$

Goal: Find an estimate $\hat{f} = \delta(X)$ of f.

Supervised learning

Learning problem:

$$Y = f(X)$$

Goal: Find an estimate $\hat{f} = \delta(X)$ of f.

For supervised learning, we are given labeled data

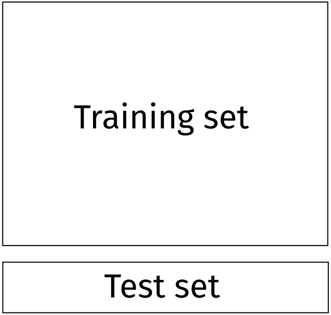
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Measure the performance of $\hat{f}$ using the empirical risk:

$$R_{\text{emp}}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{f}(x_i)).$$

Model's performance

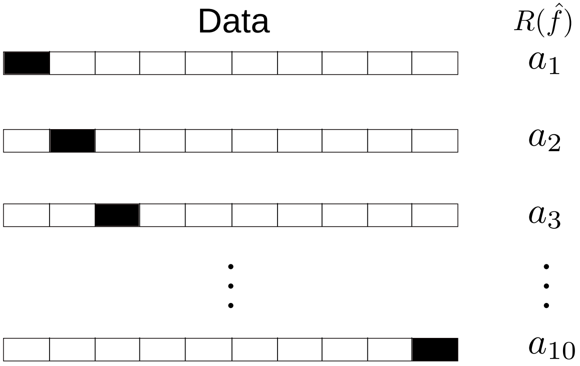
1. Train-test split



- Compute $\hat{f} = \delta(X)$ from training set X.
- Compute $R_{\text{emp}}(\hat{f})$ on the test set.

Model's performance

2. Cross-validation



CV loss: $\dfrac{a_1 + a_2 + \ldots + a_{10}}{10}$