Principles of Regularization

Prof. Dr. Josif Grabocka

Machine Learning Course
Winter Semester 2020/2021

Albert-Ludwigs-Universität Freiburg

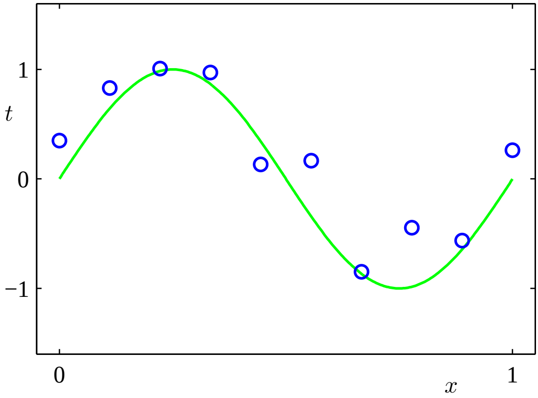
grabocka@informatik.uni-freiburg.de

Overview

Table of Contents

Example: Nonlinear regression

Consider a curve fitting example where we are given a labeled data set of N examples $\langle (x_i, y_i) \rangle_{i=1}^{N}$. Labels are generated from the target function $\sin(2\pi x)$ plus a bit of Gaussian noise.



Polynomial Regression

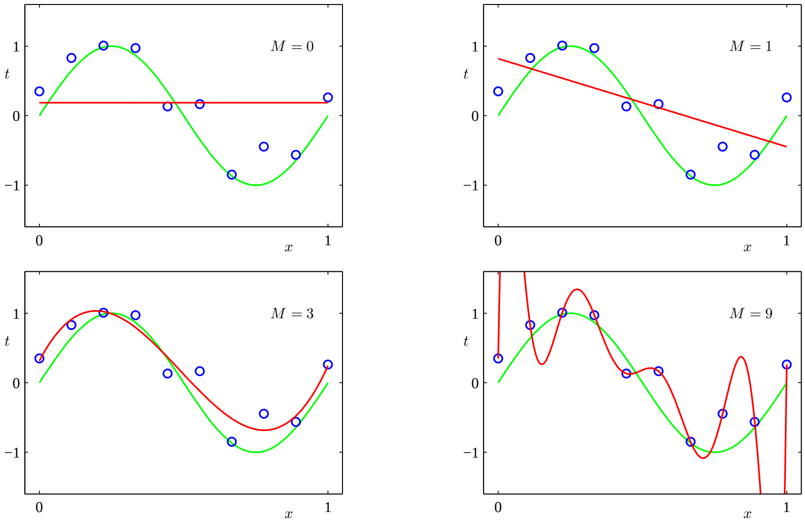
For $x \in \mathbb{R}^1$ the polynomial prediction model of degree M is:

$$\hat{y} = f(x, \theta) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \ldots + \theta_M x^M = \sum_{j=0}^{M} \theta_j x^j$$

The optimal θ^* are learned by minimizing the empirical risk:

$$\theta^* := \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} (f(x_i, \theta) - y_i)^2$$

How to choose *M*?



Generalization Performance

- Overfitting: Model perfectly fits the training data (incl. noise)
- Underfitting: Model fails to fit the training data
- Generalization: Model is accurate on test data

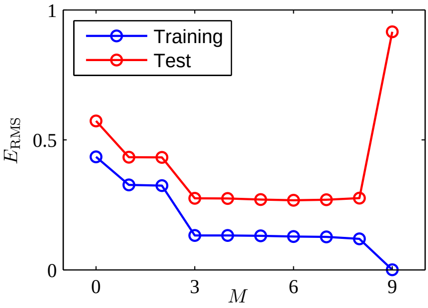


Table of Contents

Preliminary - Expected Value

z is a **continuous** r.v. $\rightsquigarrow$ $$E[z] = \int_z z\ p(z)\ dz$$

z is a c.r.v. **conditional** to v $\rightsquigarrow$ $$E_{z|v}[z] = \int_z z\ p(z|v)\ dz$$

Preliminary - Expected Value

z is a **continuous** r.v. $\rightsquigarrow$ $$E[z] = \int_z z\ p(z)\ dz$$

z is a c.r.v. **conditional** to v $\rightsquigarrow$ $$E_{z|v}[z] = \int_z z\ p(z|v)\ dz$$

Table 1: Gender (x) and Height (y)

x	y
m	180
m	170
f	160
f	170
m	170
f	160
m	170

z is a **discrete** r.v. $\rightsquigarrow$

$$E[z] = \sum_z z\ p(z) \quad E_{z|v}[z] = \sum_z z\ p(z|v)$$

$$E[y] = 160\frac{2}{7} + 170\frac{4}{7} + 180\frac{1}{7} = 168.6$$

$$E_{y|x=f}[y] = 160\frac{2}{3} + 170\frac{1}{3} + 180\frac{0}{3} = 163.3$$

Preliminary - Properties of Expectation

- Constant $\alpha \in \mathbb{R}$:

$$E[\alpha\ z] = \int_z \alpha\ z\ p(z)\ dz = \alpha \int_z z\ p(z)\ dz = \alpha\ E[z]$$

- Linearity of expectation:

$$E[\alpha z + \beta v] = \alpha E[z] + \beta E[v]$$

- Expectation of two uncorrelated r.v.:

$$E_{z,v}[z\ v] = E_z[E_v[z\ v]] = E_z[z\ E_v[v]] = E_v[v]\ E_z[z]$$

- Expectation of two correlated r.v.:

$$E_{z,v}[z\ v] = E_z[E_{v|z}[z\ v]] = E_z[z\ E_{v|z}[v]]$$

Expected Target

- A training dataset $D := \langle (x_i, y_i) \rangle_{i=1}^{N}$ drawn i.i.d. from a distribution $\mathcal{P}$

- Expected target of y given x

$$\bar{y}(x) = E_{y|x}[y] = \int_y y \; p(y|x) \; dy$$

Estimated Prediction Model (Regression)

- ML estimates a prediction model $\hat{f}(x,\theta)$ by minimizing:

$$E_{(x,y)\sim\mathcal{P}}\left[\left(\hat{f}(x,\theta)-y\right)^2\right]=\int_x\int_y\left(\hat{f}(x,\theta)-y\right)^2\ p(x,y)\ dxdy$$

- As $p(x,y)$ is typically unknown we approximate the error:

$$E_{(x,y)\sim\mathcal{P}}\left[\left(\hat{f}(x,\theta)-y\right)^2\right]\approx\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(x_i,\theta)-y_i\right)^2$$

- The estimated model $\hat{f}$ given a training set $D:=\langle(x_i,y_i)\rangle_{i=1}^{N}$

$$\hat{f}(x;D)=\hat{f}(x,\theta^*)\quad\text{s.t.}\quad\theta^*=\operatorname*{argmin}_{\theta}\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(x_i,\theta)-y_i\right)^2$$

Expected Test Error and Prediction Model

- Expected test error of $\hat{f}(x;D)$ **given a training set** D:

$$E_{(x,y)\sim\mathcal{P}}\left[\left(\hat{f}(x;D)-y\right)^2\right]=\int_x\int_y\left(\hat{f}(x;D)-y\right)^2\ p(x,y)\ dxdy$$

- Since $(x,y)\sim\mathcal{P}$ then $D:=\langle(x_i,y_i)\rangle_{i=1}^N\sim\mathcal{P}^N$ is also a r.v.

- The expected prediction model over sampled datasets is:

$$\bar{f}(x):=E_{D\sim\mathcal{P}^N}\left[\hat{f}(x;D)\right]=\int_D\hat{f}(x;D)\ p(D)\ dD$$

Expected Test Error

- Expected test error of $\hat{f}(x; D)$ **over training sets** D:

$$E_{\substack{(x,y)\sim\mathcal{P} \\ D\sim\mathcal{P}^N}}\left[\left(\hat{f}(x;D)-y\right)^2\right] = \int_x \int_y \int_D \left(\hat{f}(x;D)-y\right)^2 p(x,y)\, p(D)\, dx\, dydD$$

- The bias-variance decomposition shows how this expected test error is expressed as a sum of:
  - Bias: How well does the prediction model fits the target?
  - Variance: How much do the predictions of models varz when trained on different sampled training sets?
  - Noise: How much of the target variable cannot be unexplained?

Bias-Variance Tradeoff (I)

$$
\begin{aligned}
E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right] &= E_{x,y,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)+\bar{f}(x)-y\right)^2\right] \\
&= E_{x,y,D}\left[\left(\left(\hat{f}(x;D)-\bar{f}(x)\right)+\left(\bar{f}(x)-y\right)\right)^2\right] \\
&= E_{x,y,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right. \\
&\qquad + \left(\bar{f}(x)-y\right)^2 \\
&\qquad \left. +2\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]
\end{aligned}
$$

- Leads to:

$$
\begin{aligned}
E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right] =& E_{x,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right]+E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] \\
&+2E_{x,y,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]
\end{aligned}
$$

Bias-Variance Tradeoff (II)

$$E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right] = E_{x,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right] + E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] + \underbrace{2E_{x,y,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]}_{\text{Cancels out}}$$

$$\begin{aligned}
&E_{x,y}\left[E_D\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]\right] \\
&= E_{x,y}\left[\left(\bar{f}(x)-y\right)\left(E_D\left[\hat{f}(x;D)-\bar{f}(x)\right]\right)\right] \\
&= E_{x,y}\left[\left(\bar{f}(x)-y\right)\left(\bar{f}(x)-\bar{f}(x)\right)\right] = 0
\end{aligned}$$

Bias-Variance Tradeoff (II)

$$E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right] = E_{x,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right] + E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] + \underbrace{2E_{x,y,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]}_{\text{Cancels out}}$$

$$\begin{aligned} & E_{x,y}\left[E_D\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y\right)\right]\right] \\ &= E_{x,y}\left[\left(\bar{f}(x)-y\right)\left(E_D\left[\hat{f}(x;D)-\bar{f}(x)\right]\right)\right] \\ &= E_{x,y}\left[\left(\bar{f}(x)-y\right)\left(\bar{f}(x)-\bar{f}(x)\right)\right] = 0 \end{aligned}$$

- So far we decomposed the estimated test error to:

$$E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right] = \underbrace{E_{x,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right]}_{\text{Variance}} + \underbrace{E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right]}_{\text{Reduce further ...}}$$

Bias-Variance Tradeoff (III)

$$
\begin{aligned}
E_{x,y}\left[\left(\bar{f}(x)-y\right)^2\right] =& E_{x,y}\left[\left(\bar{f}(x)-\bar{y}(x)+\bar{y}(x)-y\right)^2\right] \\
=& E_{x,y}\left[\left(\bar{f}(x)-\bar{y}(x)\right)^2\right] \\
&+ E_{x,y}\left[\left(\bar{y}(x)-y\right)^2\right] \\
&+ \underbrace{E_{x,y}\left[2\left(\bar{f}(x)-\bar{y}(x)\right)\left(\bar{y}(x)-y\right)\right]}_{\text{cancels out}}
\end{aligned}
$$

$$
\begin{aligned}
E_{x,y}\left[\left(\bar{f}(x)-\bar{y}(x)\right)\left(\bar{y}(x)-y\right)\right] &= E_x\left[E_{y|x}\left[\left(\bar{f}(x)-\bar{y}(x)\right)\left(\bar{y}(x)-y\right)\right]\right] \\
&= E_x\left[\left(\bar{f}(x)-\bar{y}(x)\right)E_{y|x}\left[\left(\bar{y}(x)-y\right)\right]\right]
\end{aligned}
$$

This term cancels out:

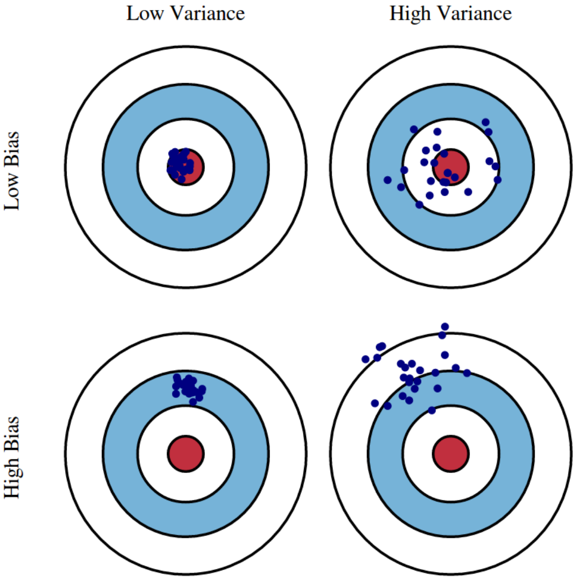
$$
E_{y|x}\left[\bar{y}(x)-y\right] = E_{y|x}\left[\bar{y}(x)\right] - E_{y|x}\left[y\right] = \bar{y}(x)-\bar{y}(x) = 0
$$

Bias-Variance Tradeoff (Finale)

$$E_{x,y,D}\left[\left(\hat{f}(x;D)-y\right)^2\right]=$$

$$\underbrace{E_{x,D}\left[\left(\hat{f}(x;D)-\bar{f}(x)\right)^2\right]}_{\text{Variance}}+\underbrace{E_{x,y}\left[\left(\bar{f}(x)-\bar{y}(x)\right)^2\right]}_{\text{Bias}^2}+\underbrace{E_{x,y}\left[\left(\bar{y}(x)-y\right)^2\right]}_{\text{Noise}}$$

Darts Example



Illustration

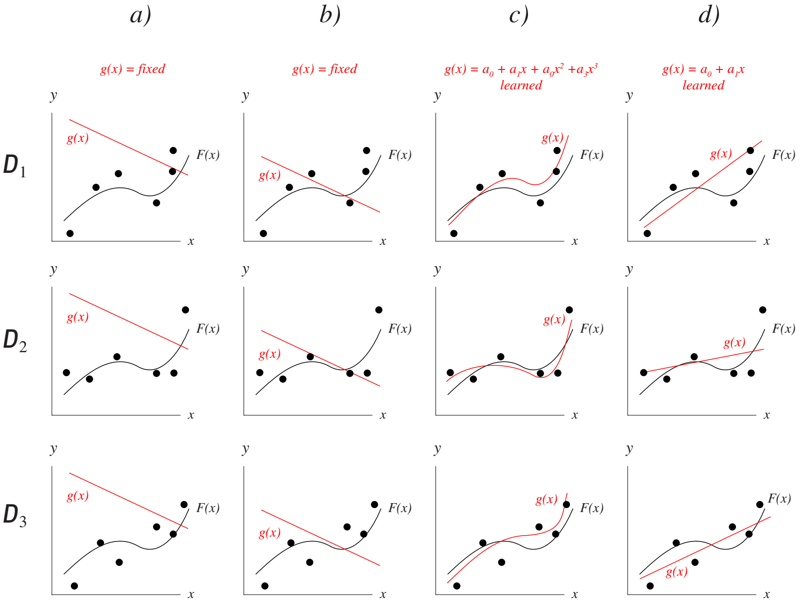
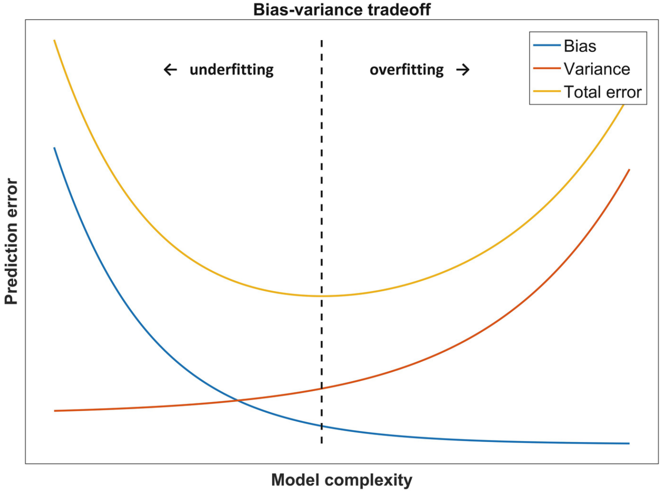


Table of Contents

Model complexity matters



Source: Dankers et al., 2018

Interpretation

- The choice of your model complexity will either:
  - increase the bias $\rightsquigarrow$ decrease the variance;
  - increase the variance $\rightsquigarrow$ decrease the bias.
- You cannot not control the noise.

Table 2: Understanding Variance, Bias, Noise

Term	High	Low
Variance	Risk of overfitting	Generalization
Bias	Risk of underfitting	Fitting
Noise	Challenging Task	Easy Task

Find a model complexity that has both a **Low Bias** (able to fit well) and a **Low Variance** (able to generalize).

Weight Decay - L1/L2 Regularization

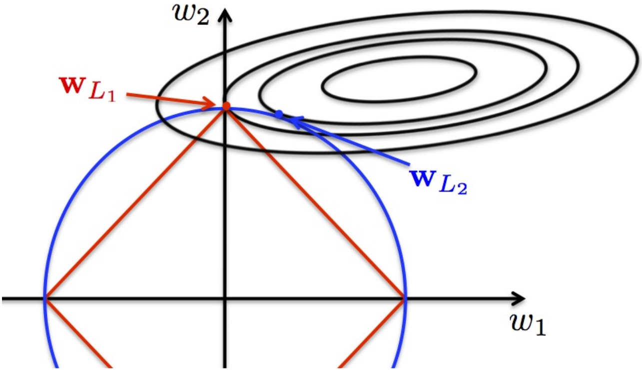
- Add a penalty term to the empirical risk ($\alpha \in \mathbb{R}_+$):

$$\underset{\theta}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f\left(x_i; \theta\right)\right) \right] + \alpha \Omega(\theta)$$

- The regularization penalizes high parameter values:

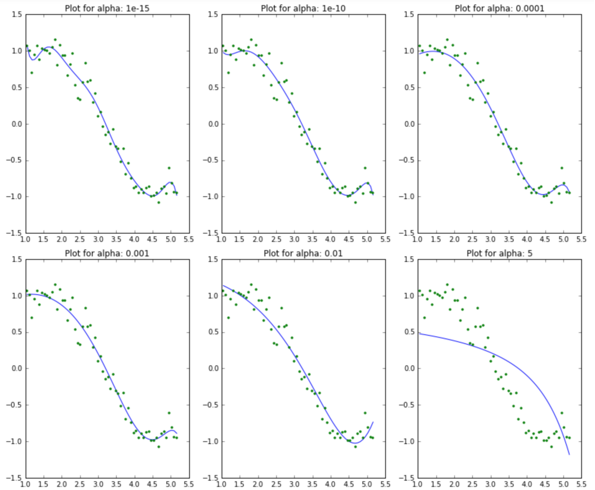
L1: $\Omega(\theta) = \frac{1}{|\theta|} \sum_{k=1}^{|\theta|} |\theta_k|$ L2: $\Omega(\theta) = \frac{1}{|\theta|} \sum_{k=1}^{|\theta|} \theta_k^2$

Illustration of the L1/L2 Regularizations



Source: `g2pi.tsc.uc3m.es`

Regularizing a Polynomial Regression



$$f(x,\theta) = \sum_{j=0}^{15} \theta_j x^j$$ (Source: `www.analyticsvidhya.com`)

Take-Home Recipe

- Bias-variance Tradeoff is a fundamental concept in ML
- Low bias and Low variance models are demanded
- Models should be regularized when exhibiting High Variance
    - Do not over-regularize to the point of having a very high bias
    - Remember the aim is accurate generalization