

ML. 1. General intro and intro to supervised learning

DataLab ICMAT

Objectives and schedule

A broad overview of Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Contents

- Key concepts: Training, testing, validating,...
- Key conceptual strategies: MLE, MLE+regularisation, Bayesian
- Key data analytic strategies

Machine learning

From Wikipedia

ML: the study of computer algos that improve automatically through experience. It is seen as a part of AI. ML algos build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so....

A subset of ML is closely related to **computational statistics**, which focuses on making predictions using computers; but not all ML is statistical learning.

The study of **mathematical optimization** delivers methods, theory and application domains to the field of machine learning.

Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.

In its application across business problems, machine learning is also referred to as predictive analytics.

Some ML examples. Red matters!!!

Uncertainty is almost ubiquitous in ML:

- Given a certain transaction, is it fraudulent or not? If fraudulent should I stop it?
- Given the monitoring trace of an Inet device, are we facing an attack? Should I stop operations?
- Does this medical image correspond to a person with a certain illness? Should I make further tests?
- A person with these FB likes will buy this type of beer? Should I send him my brand add?
- A person with these tweets is conservative? Should I send her my party propaganda?
- Robots (or ADS): If robot performs this, How will the user react? And the environment? Consequently, what should the robot do?

In applications, we'll need to go beyond

- Beyond a model with good fit...
- Beyond a model that predicts well...

In applications, we'll need to go beyond

- Beyond a model with good fit...
- Beyond a model that predicts well...
- Fraud detection. Classification problem
 - Few false positives. FPR
 - Few false negatives. FNR

In applications, we'll need to go beyond

- Beyond a model with good fit...
- Beyond a model that predicts well...
- Fraud detection. Classification problem
 - Few false positives. FPR
 - Few false negatives. FNR
 - But what really matters are minimising monetary losses!!!

In applications, we'll need to go beyond

- Beyond a model with good fit...
- Beyond a model that predicts well...
- Fraud detection. Classification problem
 - Few false positives. FPR
 - Few false negatives. FNR
 - But what really matters are minimising monetary losses!!!
- Reservoir system management. Forecasting model for inputs and demands
Feeds decision model e.g to minimize energy deficit, wasted water, given constraints.....
- Aviation safety risk management. Forecasting models for accidents and incidents, as well as their multiple impacts
Feed a risk management model: optimal safety resource allocation given constraints...
- Robot control. Forecasting model for user and environments
Feeds robot control model: optimal robot decisions over time, given constraints...

ML, Stats in modern times (Big Data)

Computer Age Statistical Inference

Volume. Space scalability

Variety. Text, images, sound, video,... video,....

Velocity. High frequency for data and decisions, time series, dynamic models. Time scalability

Create **v**alue by actually supporting decisions!!!

Machine learning



Machine learning

A computer program learns from experience E with respect to class T of tasks and performance measure P

if

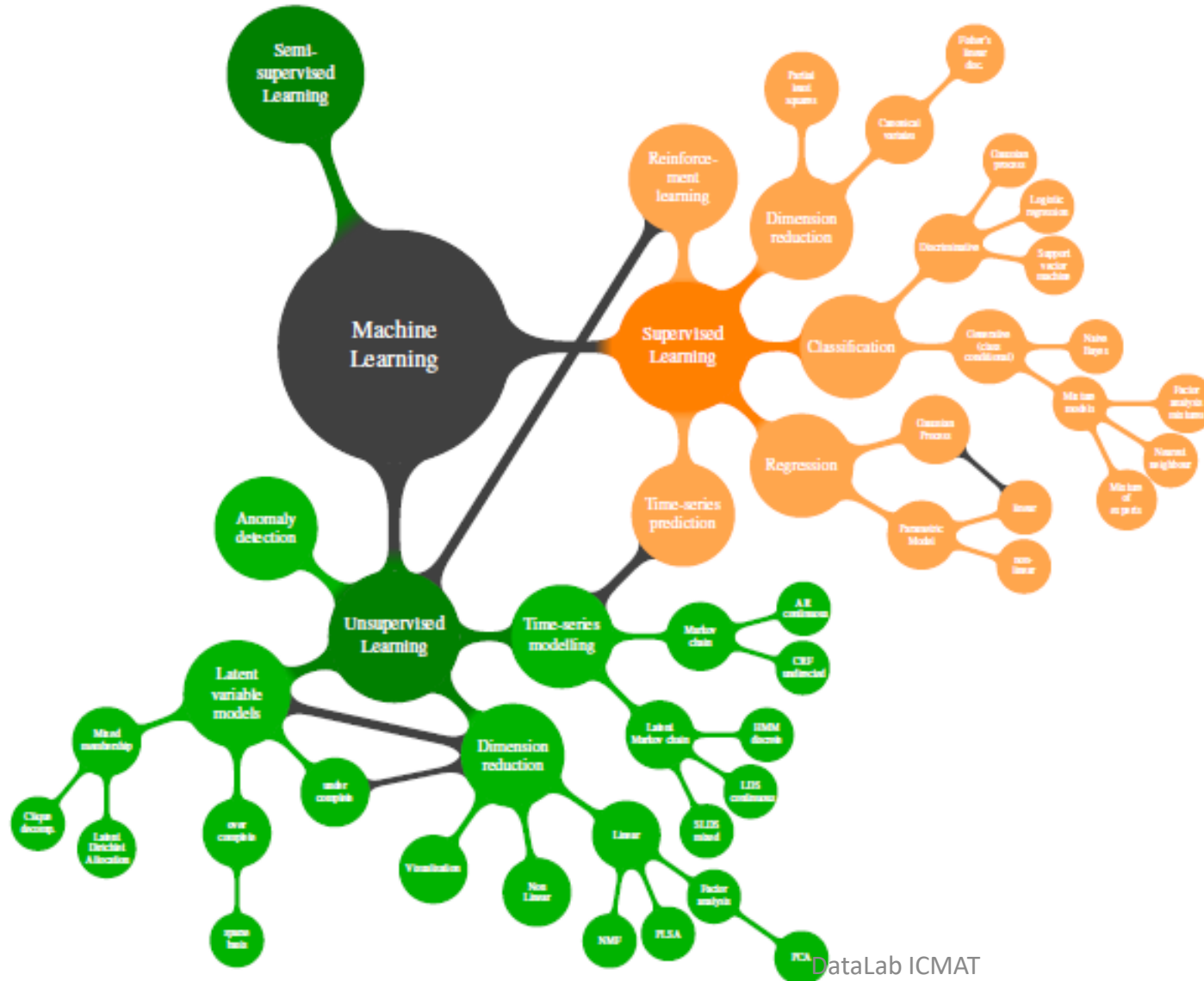
Its behaviour with respect to task T , measured according to P , improves with experience E

Representation-Evaluation-Optimization

Core themes

- **Supervised learning:** Pairs input-output available
Regression, Classification
Case: Cannabinoid detection
- **Unsupervised learning:** Outputs not available (or the inputs are the outputs)
Density estimation, clustering, outlier detection, Visualisation
Case: Singular community detection
- **Reinforcement learning:** Decisions impacting outputs on-the-fly
Markov decision processes
Case: Autonomous driving systems

With somewhat blurry borders....



Case: Lookalike modeling

Time for a KitKat



VS



A basic example

Consider a cyberattack recovery protocol for SMEs. Introduce a process.
Want to assess it. E.g to compare it with another

Test it against 12 attacks, effective 9 times (e.g., system up in less than 1 hour)

A basic example. Model

- Number X of successes in n trials (ii)
- Success probability in one trial
- Distribution of number of successes
- For $X=9$,

$$\begin{aligned}\theta_1 \\ X | \theta_1 &\sim \text{Bin}(12, \theta_1) \\ P(X=9 | \theta_1) &\propto \theta_1^9 (1-\theta_1)^3, \quad \theta_1 \in [0, 1]\end{aligned}$$

A basic example. MLE

Likelihood

Log-likelihood

Maximise likelihood or maximize log likelihood . MLE

In this case, MLE is

Defects?

For future observations (e.g. 4 successes in next 7 trials)

$$l(\theta_1) \propto \theta_1^9 (1-\theta_1)^3$$

$$h(\theta_1) = \log(l(\theta_1)) = 9 \log \theta_1 + 3 \log (1-\theta_1)$$

$$h'(\theta_1) = 0 \Rightarrow \hat{\theta}_1 = \frac{9}{12} = .75$$

$$Pr(Y=4 | \hat{\theta}_1, 7) = \binom{7}{4} .75^4 .25^3$$

A basic example. Bayes

Prior, e.g.

$$\pi(\theta_1) = 1$$

$$\pi(\theta_1|y) \propto 1 \times \theta_1^9 (1-\theta_1)^3 \sim \text{Be}(10, 4)$$

Posterior

Posterior mean

$$\frac{10}{14}$$

Posterior mode (MAP)

$$\frac{9}{12}$$

Predictive

$$\begin{aligned} P(Y=4|y) &= \int \binom{7}{4} \theta_1^4 (1-\theta_1)^3 \pi(\theta_1|y) d\theta_1 \\ &= \frac{\binom{7}{4} \binom{13}{3}}{\binom{20}{12}} \end{aligned}$$

And so...

We end up using this to make decisions. Which protocol to implement?

How would you choose between two protocols?

Intro to Supervised Learning

SL: ingredients

- Data available: examples, samples, instances,...
- Several observed variables: predictors, attributes, features, covariates, explanatory variables, independent variables,...
- Some of special interest: response(s), dependent variable(s), target(s), output(s), label(s),...

SL: types of problems

1. Regression, response variable is continuous
2. Classification, response variable is discrete
3. Other:
 - Mixed (some continuous, some discrete)
 - Discrete but ordered
 - ...

Predictors

(x_1, \dots, x_p)

Dependent variable

y

Some relation

$$y = f(x) + \epsilon$$

Systematic info

Random term. Zero mean, Indep of x

SL: objectives

Prediction. Predict the response for new observations

$$\hat{y} = \hat{f}(x)$$

Black box. Don't care about form, except that predicts nicely.....

Accuracy. Reducible and Irreducible

$$\begin{aligned} E(y - \hat{y})^2 &= E(f(x) + \varepsilon - \hat{f}(x))^2 = \\ &= \underbrace{(f(x) - \hat{f}(x))^2}_{\text{RED.}} + \underbrace{\text{Var}(\varepsilon)}_{\text{IRRED.}} \end{aligned}$$

SL: objectives

Inference. Obtain information about relation between independent variables and output.

Understand the way predictors affect output

Not a black box

SL: objectives

Decision. Support decisions

Remember

How do we estimate f ?

Training data

$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$

$$\hat{f}(x_i) \approx y_i$$

For any observable

$$\hat{f}(x_0) \approx y_0$$

Parametric, e.g.

$$f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

$$y \approx \beta_0 + \dots + \beta_p x_p$$

$$f \text{ vs } (\beta_0, \dots, \beta_p)$$

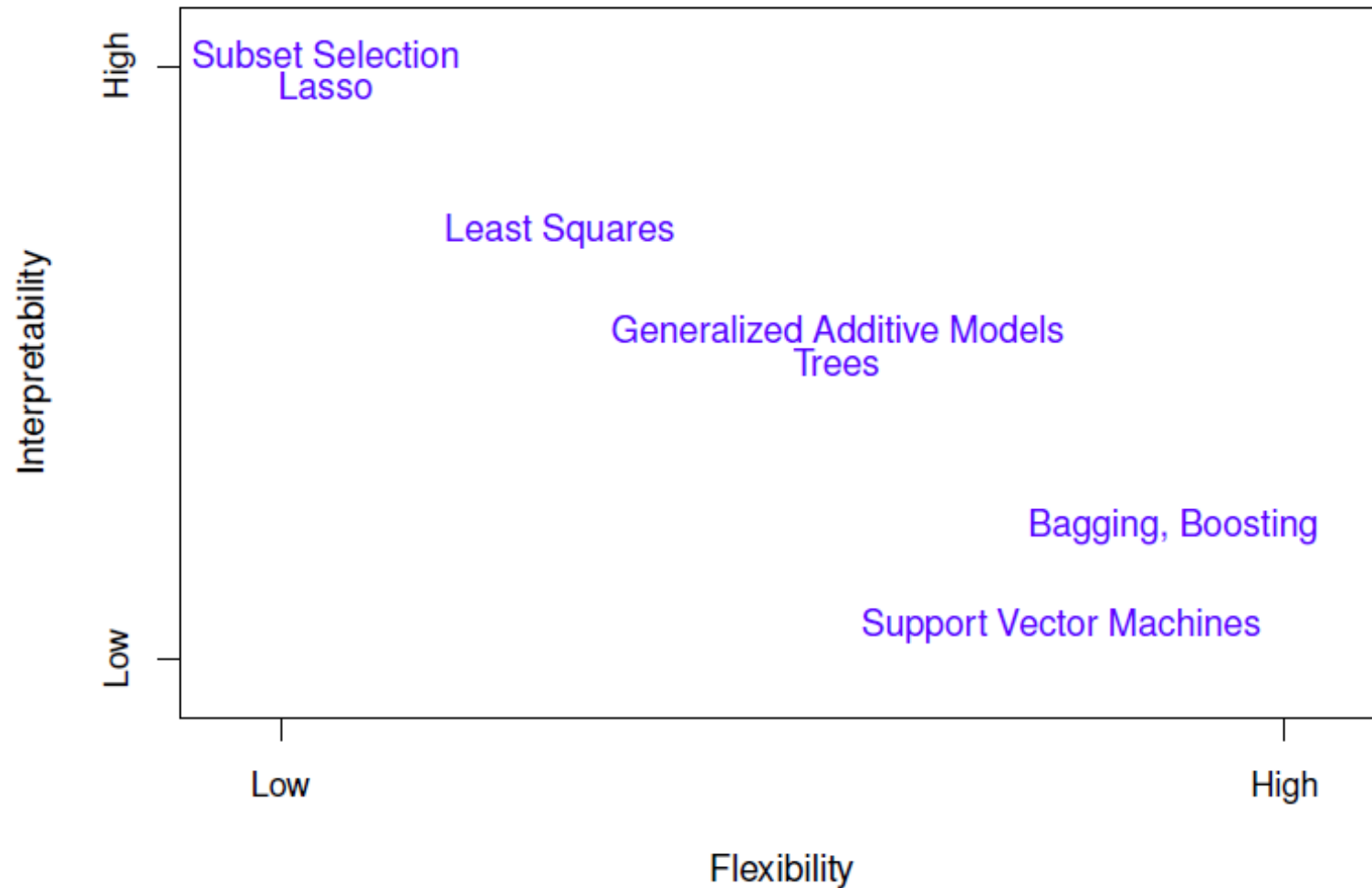
Flexibility and overfitting

Non-parametric

Wider range, much larger #observations

Flexibility vs Interpretability

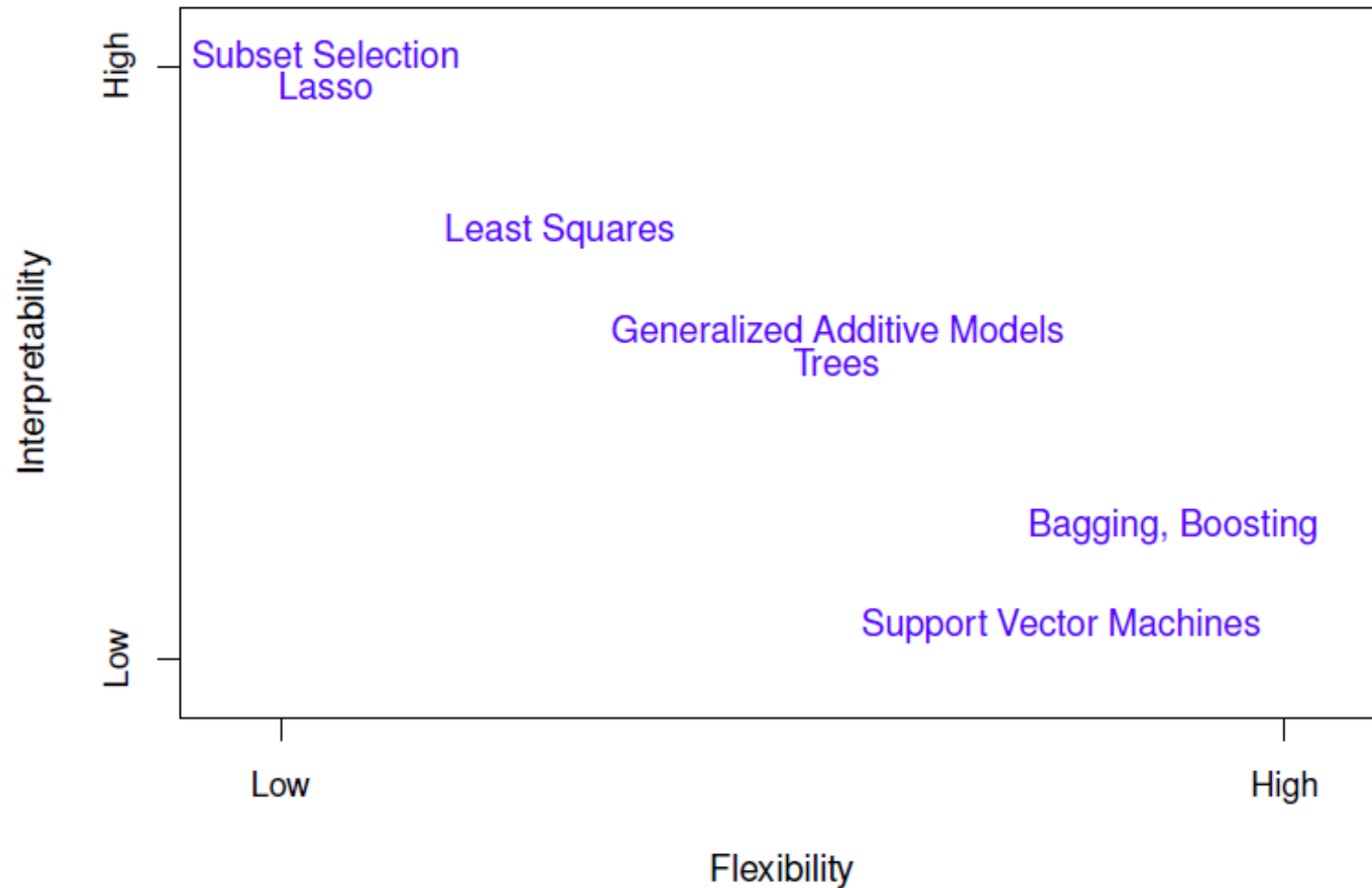
(somewhat old figure from ISLR)



Rudin's paper!!!!

Flexibility vs Interpretability

(somewhat old figure from ISLR)



Deep
Models!!!!

Rudin's paper!!!!

Assessing accuracy

No free lunches

Quality of fit, e.g.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

Training MSE vs Test MSE

Not

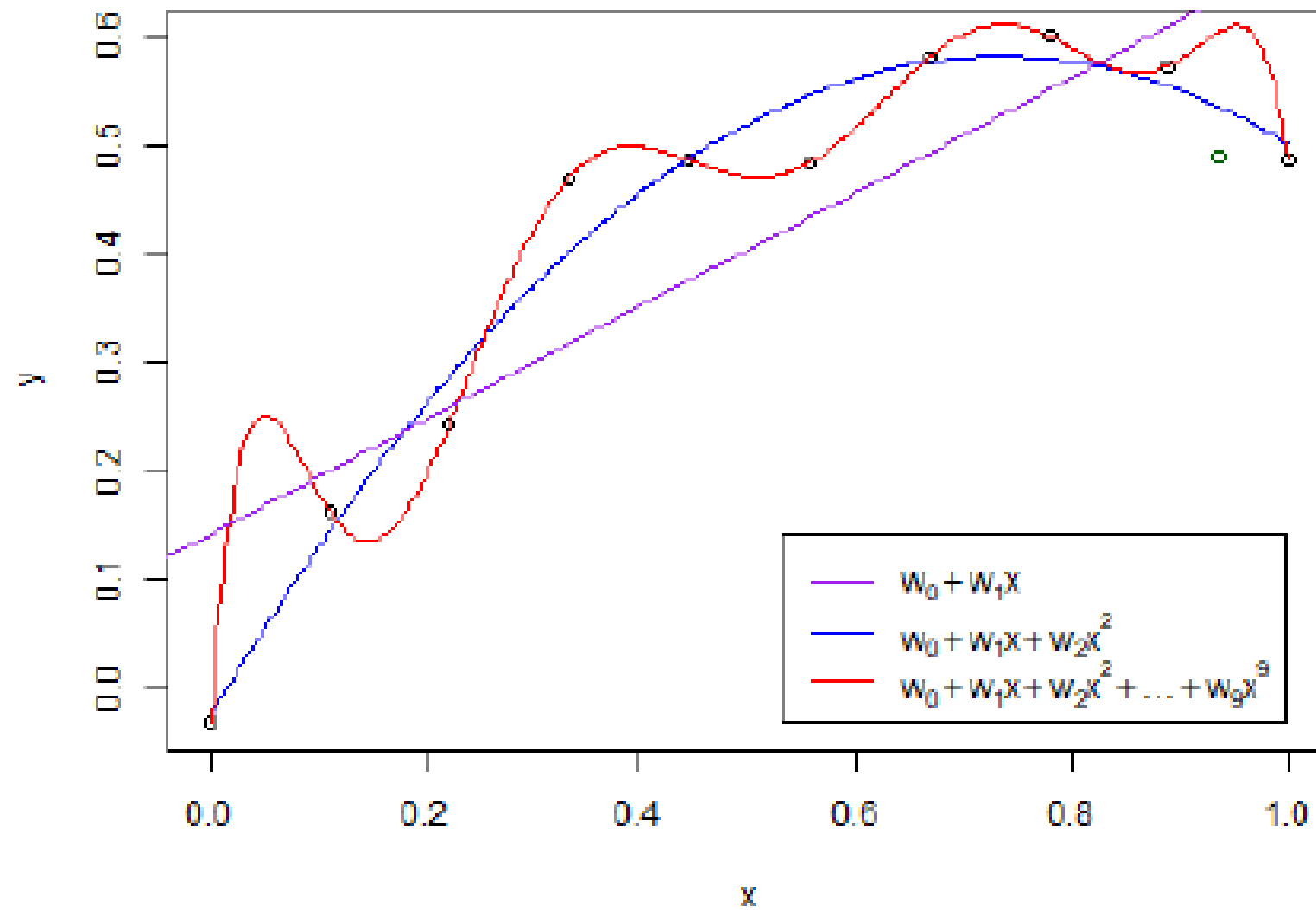
$$\hat{f}(x_i) \approx y_i$$

but

Small

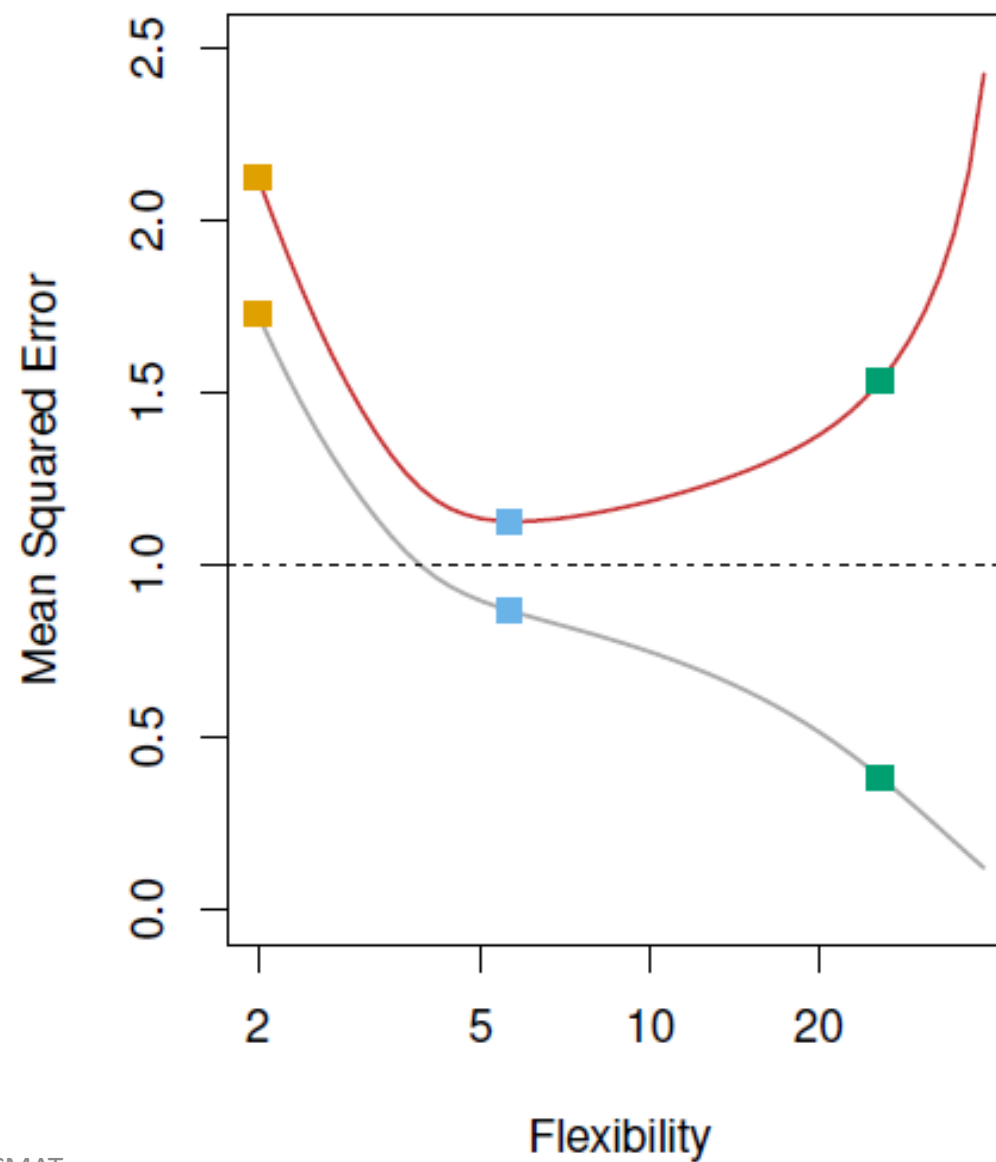
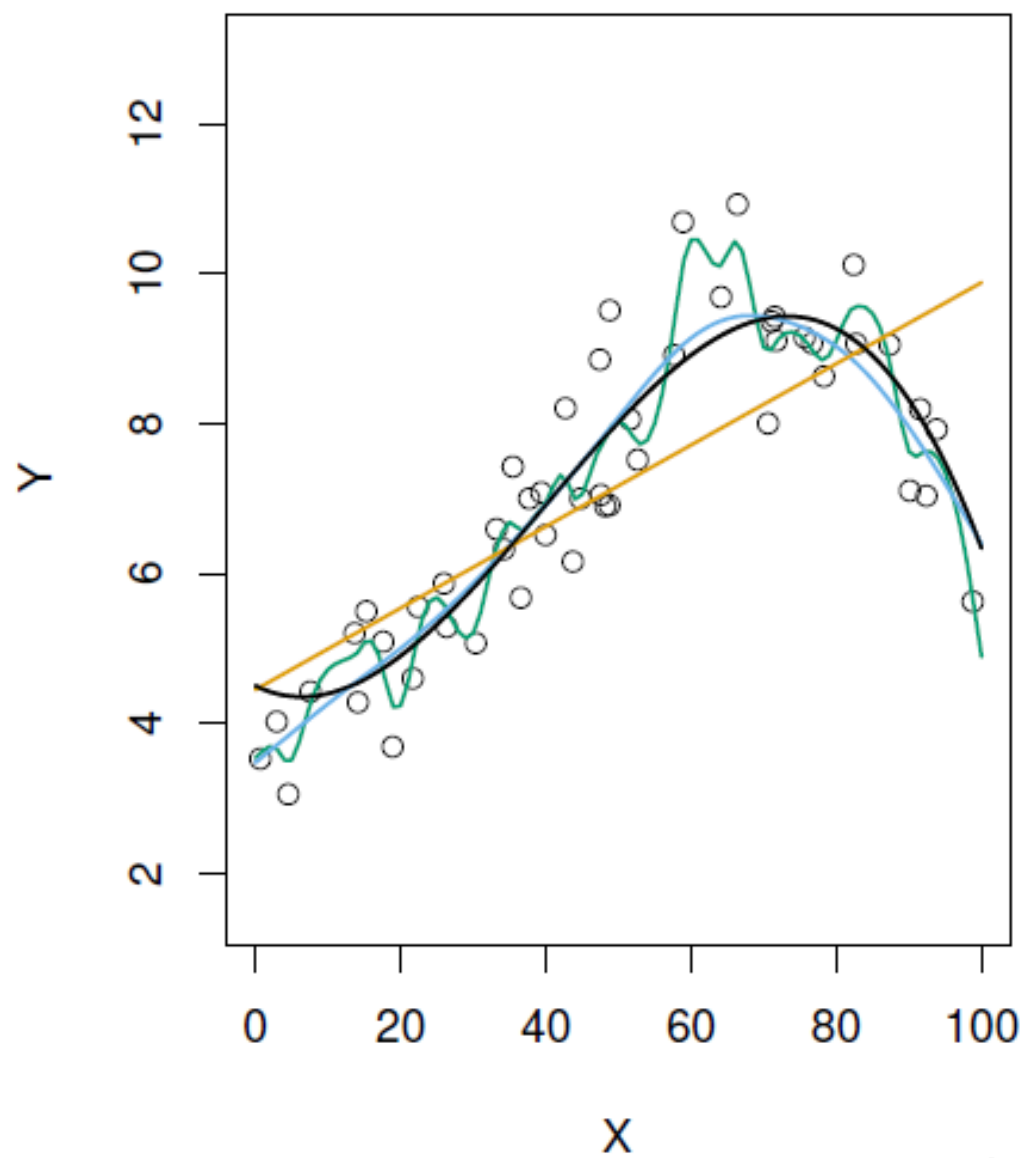
$$\hat{f}(x_0) \approx y_0$$

$$AVE (y_0 - \hat{f}(x_0))^2$$



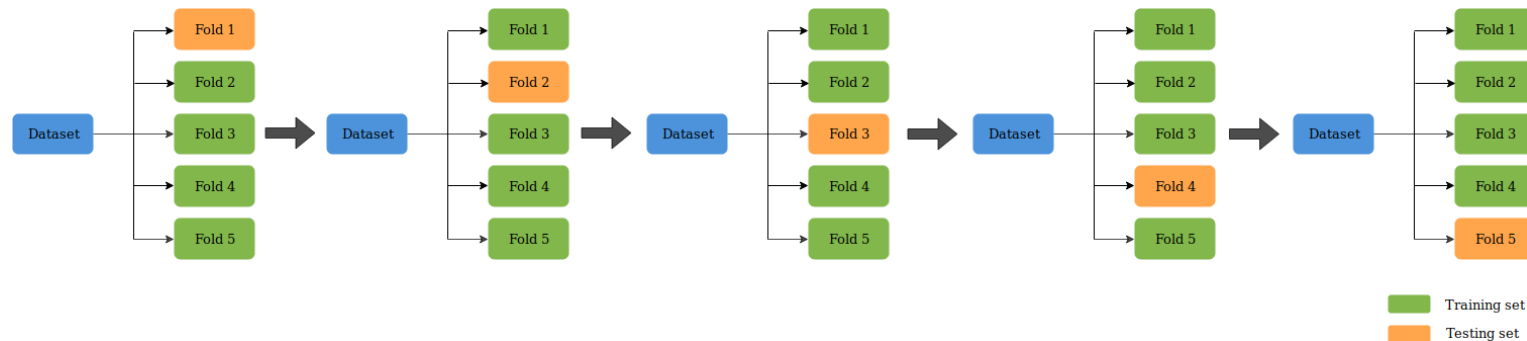
Model selection

- Compute empirical risk over training set
- May be reduced almost arbitrarily increasing model complexity
e.g. based on polynomials
- Generalisation, error over observations not used to train the model (cannabinoid project)
- If no test set available, split data in two sets:
 - Training set
 - Test set



Cross validation

- Hyperparameter choice to control model complexity
- Choose a third set for validation to select and compare models
- If data not plentiful, divide set in k partitions
- Use k-1 to train and the other to test: k models
- Cross validation error



- If $k=n$ leave-one-out cross validation

Bias-variance tradeoff

- Assume model
- Exp. Pred. error (under quad. Loss)
- Decomposed as

Variance. How approximation changes if a different training set used

Bias. Error due to using much simpler model

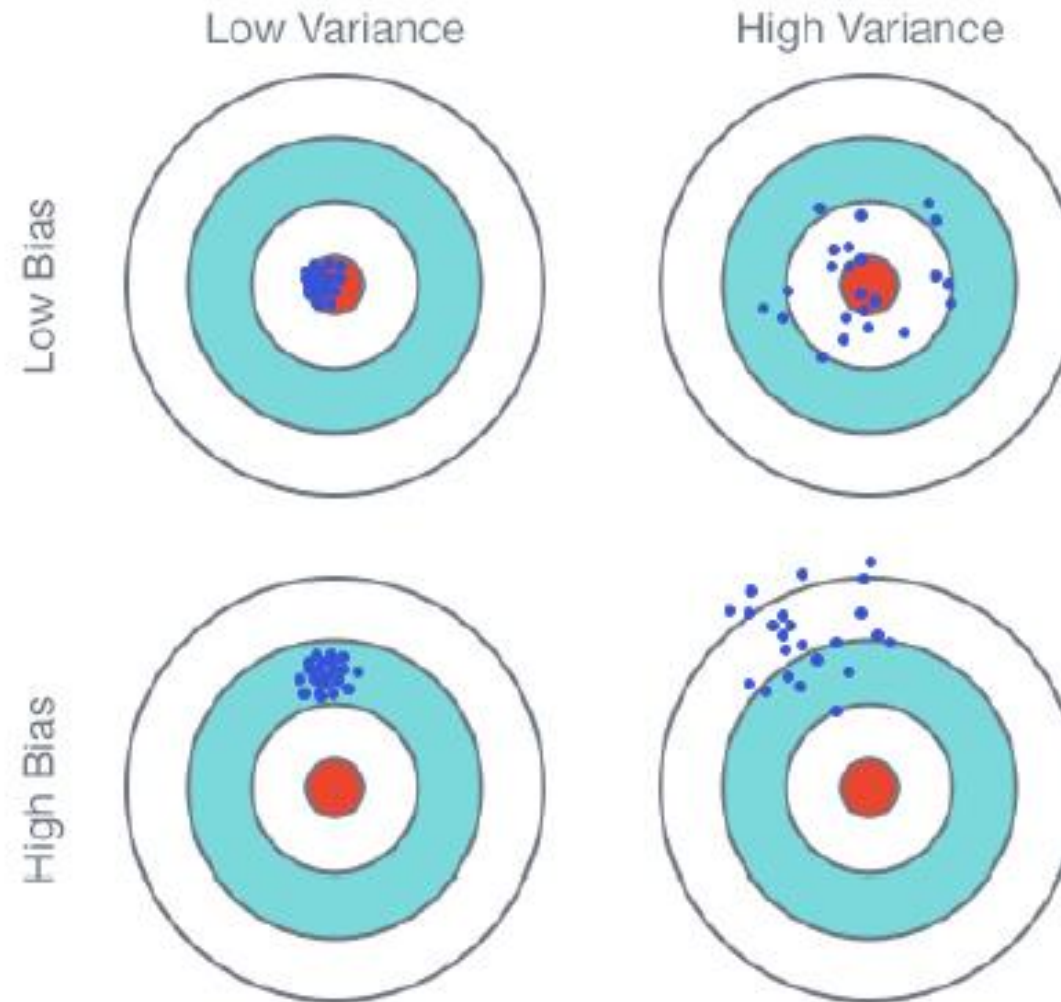
Generally, more flexible method: variance increases, bias decreases

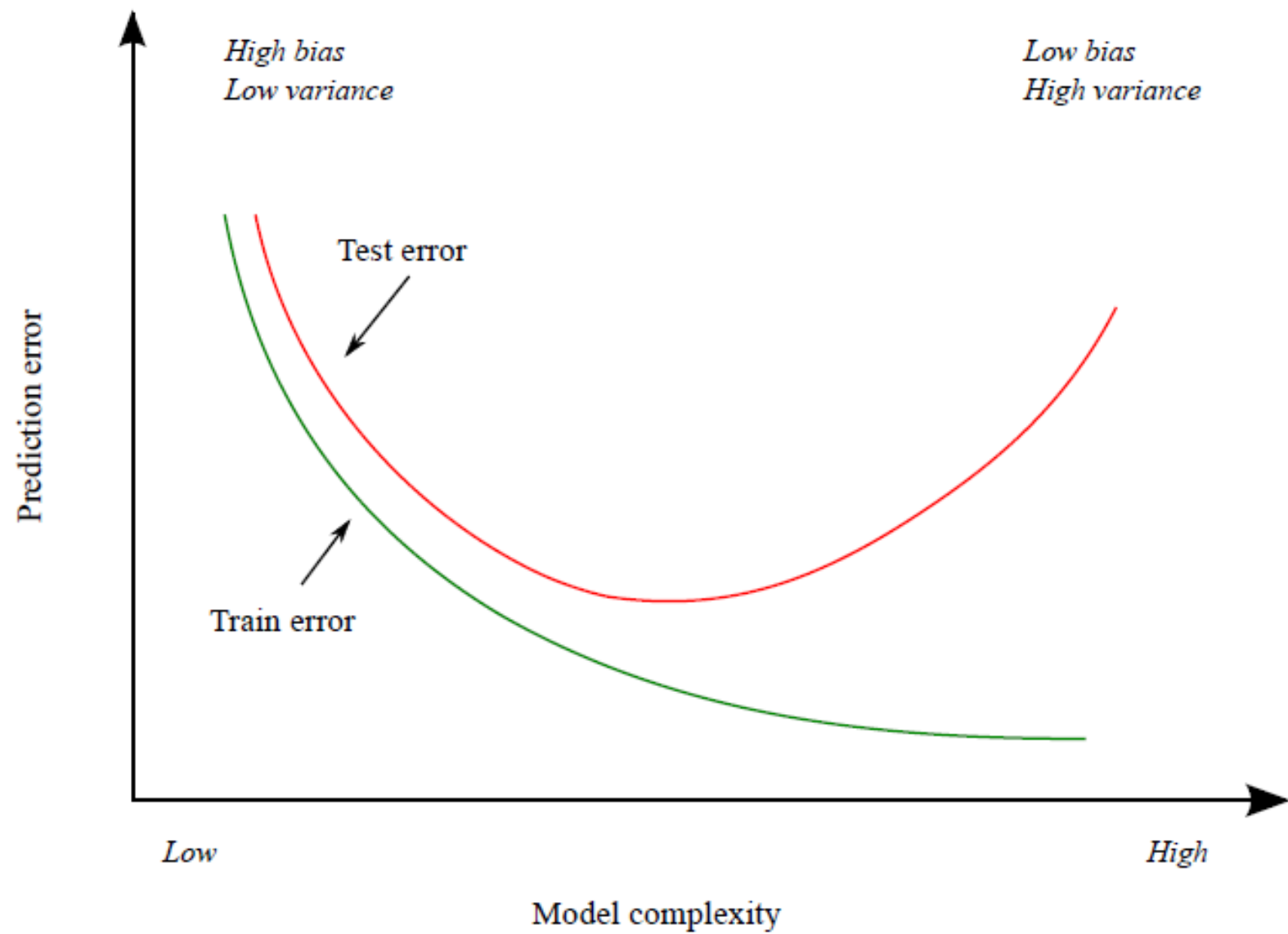
$$Y = f(X) + \varepsilon \quad \begin{array}{l} E(\varepsilon) = 0 \\ \text{Var}(\varepsilon) = \sigma^2 \end{array}$$

$$\text{EPE} = E(Y - \hat{f}(X))^2$$

$$\begin{array}{l} \text{EPE} = E(\hat{f}(X) - f(X))^2 \quad \text{BIAS}^2 \\ + \\ E(\hat{f}(X) - E(\hat{f}(X)))^2 \quad \text{VAR} \\ + \\ \sigma^2 \quad \text{NOISE} \end{array}$$

Bias-variance tradeoff





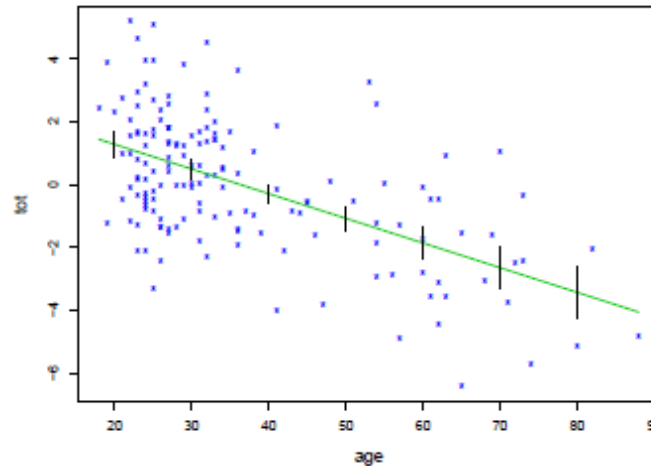
Regularisation

- Aim: reduce variance in exchange of a small bias
- Introduce sparsity
- Limit model complexity by adding a regularisation term

$$\min \sum_{i=1}^n (y_i - \beta x_i)^2 + \lambda \sum_j \beta_j^2$$

Linear regression model. A typical example

Consider a study of kidney function. Data represent (x =age of person, y =tot, a composite measure of the overall function). Kidney function declines with age. We need to provide additional information concerning decline rate. This is important in managing kidney transplants.



Numerical example in Lab. More on chap2. Here we just go through the concepts and methods.

Check

https://en.wikipedia.org/wiki/Simple_linear_regression

Example: linear regression

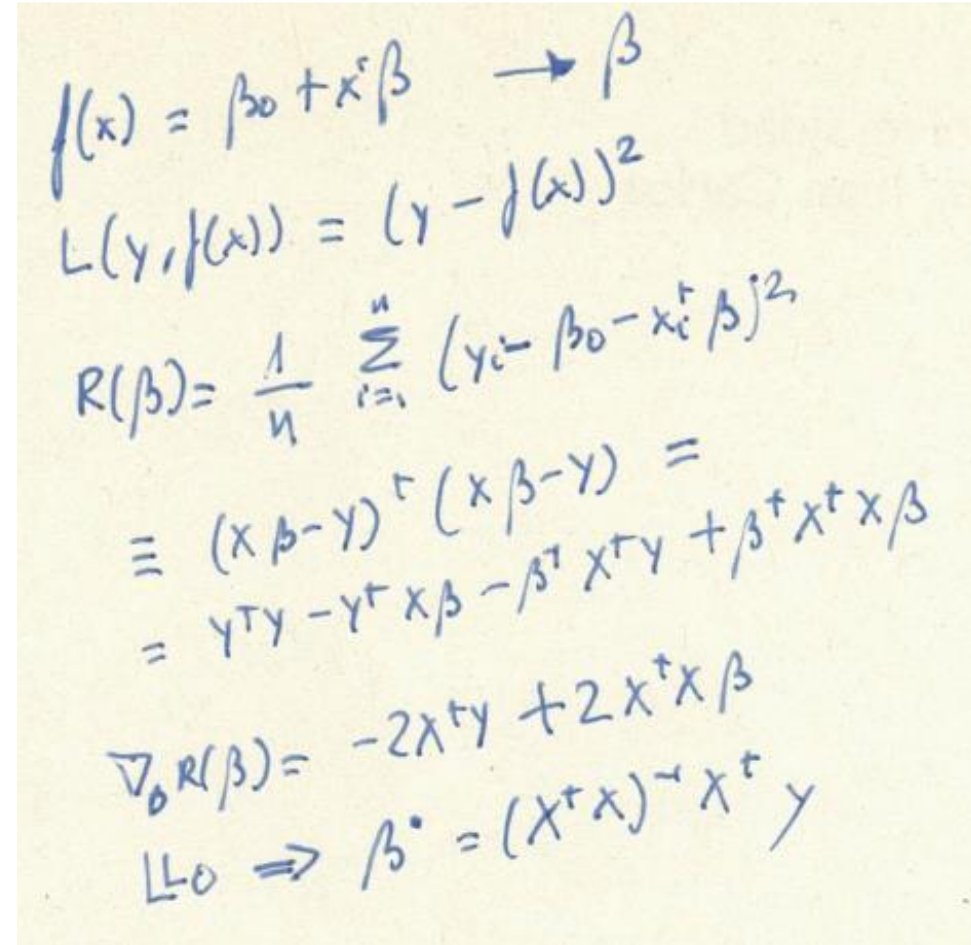
Linear functions

Quadratic loss

Empirical risk

Gradient

Minimize empirical risk



Handwritten mathematical derivations for linear regression:

$$f(x) = \beta_0 + x^T \beta \rightarrow \beta$$
$$L(y, f(x)) = (y - f(x))^2$$
$$R(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - x_i^T \beta)^2$$
$$\begin{aligned} &\equiv (X\beta - Y)^T (X\beta - Y) = \\ &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X \beta \end{aligned}$$
$$\nabla_{\beta} R(\beta) = -2X^T Y + 2X^T X \beta$$
$$L_0 \Rightarrow \beta^* = (X^T X)^{-1} X^T Y$$

Linear regression

Data structure. Response
Explanatory variables

Model

Likelihood

Log-likelihood

MLE

Handwritten mathematical derivations for linear regression:

$$Y$$
$$(x_1, \dots, x_n)$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}, \quad i = 1, \dots, n$$
$$\varepsilon_i \sim N(0, \sigma^2) \text{ IND.}$$
$$\theta = (\beta_0, \dots, \beta_p, \sigma)$$
$$p(\theta | x) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \prod_{i=1}^n \exp\left(-\frac{1}{2} \left(\frac{y_i - \beta x_i}{\sigma} \right)^2 \right)$$
$$\max -\frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \beta x_i}{\sigma} \right)^2 \dots$$
$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$s^2 = \frac{1}{(n-p)} (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

Linear regression

If n or n and p large

• COMPUTE $X = QR$ $Q_{n \times p}$ ORTH. COLUMNS, $R_{p \times p}$ UPPER TRIANG.

$$\left[(X^T X)^{-1} = (R^T Q^T Q R)^{-1} = (R^T R)^{-1} = R^{-1} (R^{-1})^T \right]$$

• COMPUTE R^{-1}

• SOLVE $R \hat{\beta} = Q^T y$

$$\left[\hat{\beta} = (X^T X)^{-1} X^T y = (R^T Q^T Q R)^{-1} R^T Q^T y \right]$$
$$= (R^T R)^{-1} R^T Q^T y = R^{-1} Q^T y$$

DataLab ICMAT

Linear regression with regulariser

If p large (much larger than n)

$$\min \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum \beta_i^2$$

$$\min \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum |\beta_i|$$

Bayesian inference with linear regression model

Model

Standard noninformative prior

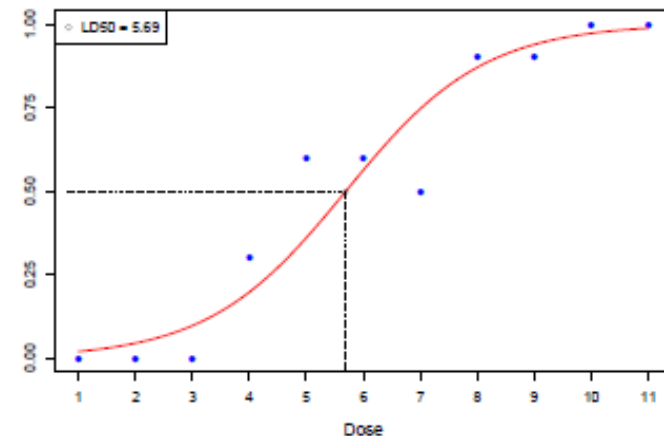
Posterior

$$\begin{aligned} y_1 &= x_1^T \beta + \varepsilon_1 \\ &\vdots \\ y_n &= x_n^T \beta + \varepsilon_n \quad \varepsilon_i \sim N(0, \sigma^2) \quad \parallel \quad y | \beta, \sigma^2, X \sim N(X\beta, \sigma^2 I) \\ p(\beta, \sigma^2) &= p(\beta | \sigma^2) p(\sigma^2) \propto \sigma^{-2} \\ p(\beta, \sigma^2 | y) &\propto \frac{p(y | \beta, \sigma^2) p(\beta, \sigma^2)}{p(y)} \\ \beta | \sigma, y &\sim N(\hat{\beta}, V_{\beta} \sigma^2) \quad V_{\beta} = (X^T X)^{-1} \\ &\quad \hat{\beta} = V_{\beta} X^T y \\ p(\sigma^2 | y) &= \frac{p(\beta, \sigma^2 | y)}{p(\beta | \sigma^2, y)} \sim \text{Inv-}\chi^2(n-p, s^2) \\ s^2 &= \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta}) \end{aligned}$$

Logistic regression. A typical example

A new anti-cancer drug is being developed. Before human testing can begin, animal studies are needed to determine safe dosages. A bioassay or dose-response experiment is carried out: 11 groups of 10 mice are treated with an increasing dose of drug and the proportion of deaths are observed.

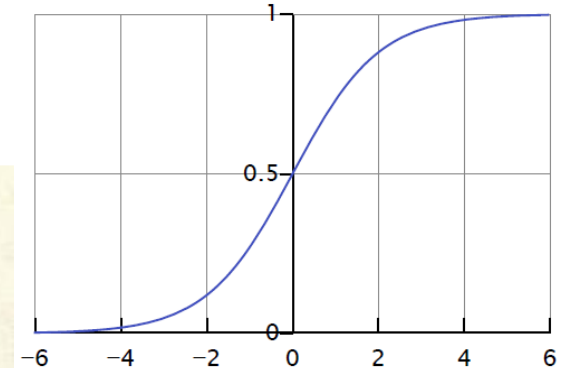
Just a quick concept note. More in Labs and chapter 3



Check

https://en.wikipedia.org/wiki/Logistic_regression

Logistic regression



Data

$(x_i, n_i, y_i) \quad i=1, \dots, K$
 x_i , EXP
 n_i , TRIALS
 y_i , SUCCESSES

Model

$$y_i | \theta_i \sim \text{Bin}(n_i, \theta_i) \quad \text{logit}(\theta_i) = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta x_i$$

$$\theta_i = \text{logit}^{-1}(\alpha + \beta x_i)$$

Likelihood

$$l(\alpha, \beta | (x, n, y)) = \prod_{i=1}^K \left[\binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i-y_i} \right]$$

$$\max_{\alpha, \beta} \sum_{i=1}^K \left[y_i \log \theta_i + (n_i - y_i) \log (1-\theta_i) \right]$$

MLE

$$\begin{aligned} &\rightarrow \hat{\alpha}, \hat{\beta} \\ &\theta_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \end{aligned}$$

Logistic regression with regulariser

$$\max \sum \left[y_i \log \text{logit}^{-1}(\alpha + \beta x_i) + (n_i - y_i) \log (1 - \text{logit}^{-1}(\alpha + \beta x_i)) \right] - \lambda \sum \beta_i^2$$

Bayesian logistic regression

Likelihood

Generic prior

Generic posterior

$$p(\alpha, \beta | y, n, X) \propto p(\alpha, \beta) \prod_{i=1}^K \left[\left(\text{logit}^{-1}(\alpha + \beta x_i) \right)^{y_i} \left(1 - \text{logit}^{-1}(\alpha + \beta x_i) \right)^{n_i - y_i} \right]$$

MCMC

Quality of classification models

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Accuracy

$$(TP+TN)/(P+N)$$

Sensitivity, recall, TPR

$$TP/(TP+FN)$$

Specificity , TNR

$$TN/(TN+FP)$$

Precision

$$TP/(TP+FP)$$

0-1 loss

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

$$Ave(I(y_0 \neq \hat{y}_0))$$

Quality of classification models

- Accuracy not sufficient in imbalanced problems
 - Example. CTR (click-through rate, click 1, no-click 0)
 - About 10^8 ads (observations), only 80000 clicks
 - A model classifying always as 0, accuracy of 99.92%
- More generally, need to take into account costs
 - Fraud, no fraud
 - Cannabinoids. Decisions entail researching or not researching drug
- Area under the (receiver operating characteristics, ROC) curve (AUC)

SL. To be seen

Regression. “Advanced topics” in linear regression. Ridge regression, Lasso, ElasticNet, Bayesian regression, Variable selection.

Classification: *KNN*, *Naive Bayes*, Bayesian classification

C+R. Decision trees, random forests, boosting

C (+R). Support vector machines

R +C. Neural networks. Deep neural nets (convolutional, recurrent, transformers)