High dimensional statistics for genomic data

Laurent Jacob

January 16, 2017

Some references

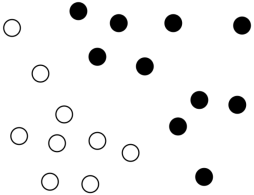
- Hastie, Tibshirani, Friedman. The Elements of Statistical Learning, 2001. (free online)
- Theoretical statistics class by P. Bartlett: http://www.stat.berkeley.edu/ bartlett/courses/2013spring-stat210b/.
- Theoretical statistics class by S. Arlot and F. Bach (in French): http://www.di.ens.fr/ arlot/2013orsay.htm.
- Boyd and Vandenberghe. Convex Optimization, 2004. (free online)
- The matrix cookbook.

Outline of this class

1. A few examples.
2. Bias/variance trade-off and how to deal with it.
3. Supervised learning.
4. Unsupervised learning.

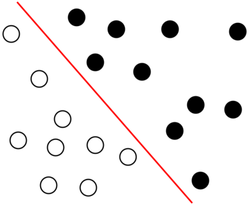
Statistical learning and applications

- This class is concerned with learning from data. Essentially:



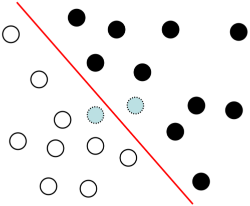
Statistical learning and applications

- This class is concerned with learning from data. Essentially:



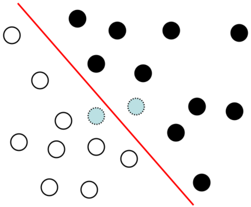
Statistical learning and applications

- This class is concerned with learning from data. Essentially:



Statistical learning and applications

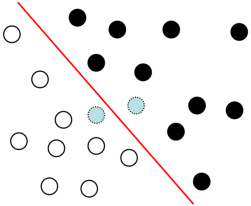
- This class is concerned with learning from data. Essentially:



- Also: multi-class, regression, unsupervised...

Statistical learning and applications

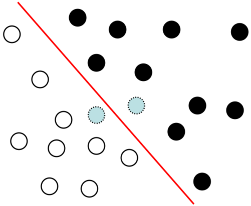
- This class is concerned with learning from data. Essentially:



- Also: multi-class, regression, unsupervised...
- We start with a few examples to make things concrete.

Statistical learning and applications

- This class is concerned with learning from data. Essentially:



- Also: multi-class, regression, unsupervised...
- We start with a few examples to make things concrete.
- These examples highlight a general problem which we will discuss right after.

Part I

A few examples

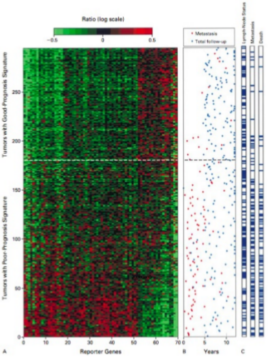
Biological data in high dimension

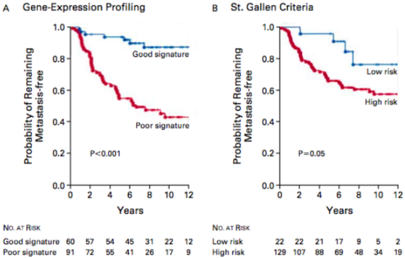
Modern technologies in molecular biology provide descriptions of individuals through thousands/millions of descriptors:

- Gene expression (arrays, sequencing),
- SNPs,
- Methylations,
- ...

Potential to allow better understanding/prediction of complex phenomena.

Tumor classification for prognosis





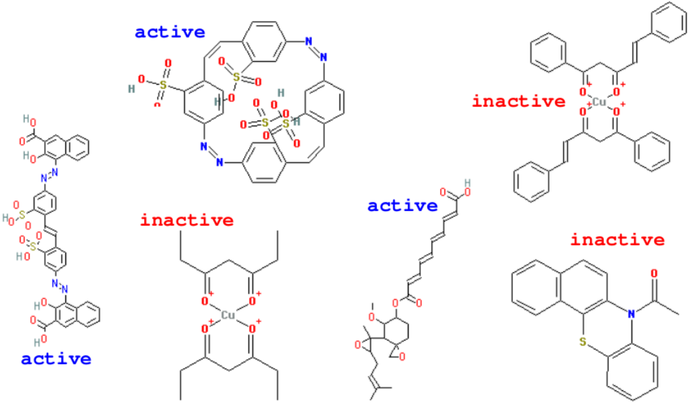
- Given the expression of the genes in a new tumor, predict the occurrence of a metastasis in the next 5 years.
- Similarly: diagnosis.

Protein classification

- Secreted proteins:
  MASKATLLLAFTLLFATCIARHQQRQQQNQCQLQNIEA...
  MARSSLFTFLCLAVFINGCLSQIEQQSPWEFQGSEVW...
  MALHTVLIMLSLLPMLEAQNPEHANITIGEPITNETLGWL... . . .
- Non secreted proteins:
  MAPPSVFAEVPQAQPVLVFKLIADFREDPDPRKVNLGVG...
  MAHTLGLTQPNSTEPHKISFTAKEIDVIEWKGDILVVG...
  MSISESYAKEIKTAFRQFTDFPIEGEQFEDFLPIIGNP... . . .

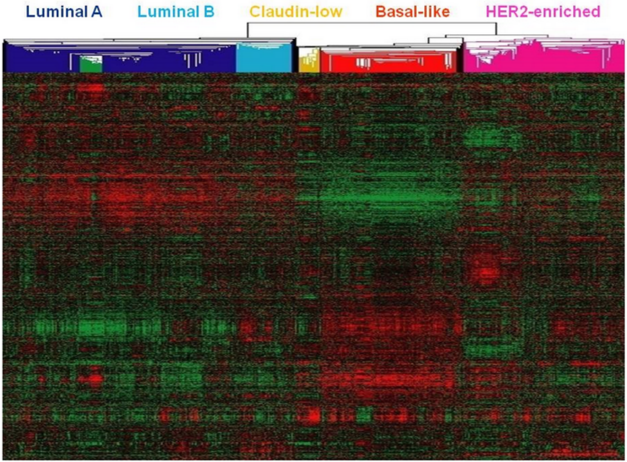
Given the sequence for a new protein, is it secreted or not?

Molecule classification for drug design



Given a candidate molecule, is it active against a therapeutical target.

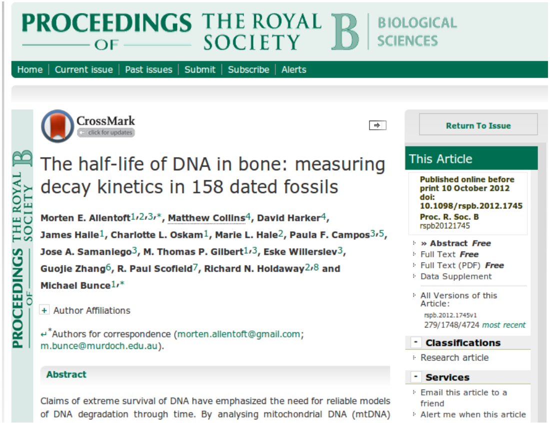
Gene expression clustering



(from C. Perou's website)

Are there groups of breast tumors with similar gene expression profile?

Ancestral genome reconstruction


PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES
Home | Current issue | Past issues | Submit | Subscribe | Alerts
CrossMark
click for updates
Return To Issue

The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils

Morten E. Allentoft[1,2,3,*], Matthew Collins[4], David Harker[4], James Haile[1], Charlotte L. Oskam[1], Marie L. Hale[2], Paula F. Campos[3,5], Jose A. Samaniego[3], M. Thomas P. Gilbert[1,3], Eske Willerslev[3], Guojie Zhang[6], R. Paul Scofield[7], Richard N. Holdaway[2,8] and Michael Bunce[1,*]

+ Author Affiliations

*Authors for correspondence (morten.allentoft@gmail.com; m.bunce@murdoch.edu.au).

Abstract

Claims of extreme survival of DNA have emphasized the need for reliable models of DNA degradation through time. By analysing mitochondrial DNA (mtDNA)

This Article

Published online before print 10 October 2012
doi: 10.1098/rspb.2012.1745
Proc. R. Soc. B
rspb20121745

- » Abstract *Free*
- Full Text *Free*
- Full Text (PDF) *Free*
- Data Supplement

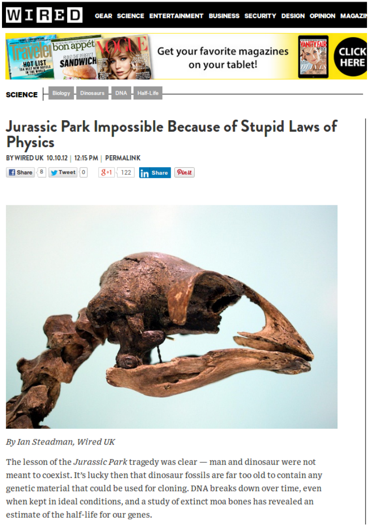
- All Versions of this Article:
  rspb.2012.1745v1
  279/1748/4724 *most recent*

- Classifications
- Research article

- Services
- Email this article to a friend
- Alert me when this article

Decay of DNA molecules.

Ancestral genome reconstruction

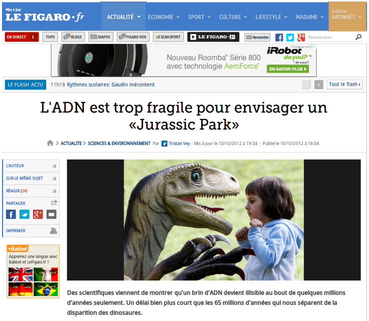

WIRED
GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION MAGAZI
Get your favorite magazines on your tablet!
CLICK HERE
SCIENCE
Biology
Dinosaurs
DNA
Half-Life
Share 8
Tweet 0
+1 122
Share
Pin it

Jurassic Park Impossible Because of Stupid Laws of Physics

BY WIRED UK 10.10.12 | 12:15 PM | PERMALINK

By Ian Steadman, Wired UK

The lesson of the *Jurassic Park* tragedy was clear — man and dinosaur were not meant to coexist. It's lucky then that dinosaur fossils are far too old to contain any genetic material that could be used for cloning. DNA breaks down over time, even when kept in ideal conditions, and a study of extinct moa bones has revealed an estimate of the half-life for our genes.



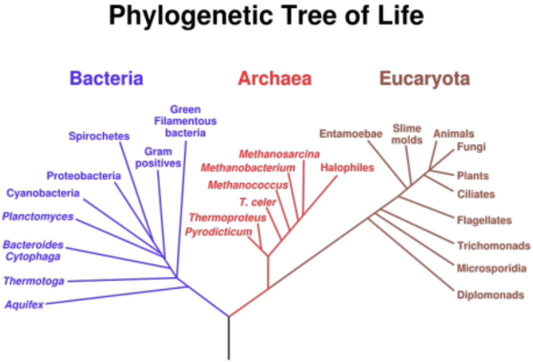
L'ADN est trop fragile pour envisager un «Jurassic Park»

ACTUALITE > SCIENCES & ENVIRONNNEMENT Par Tristan Vey | Mis à jour le 10/10/2012 à 19:34 | Publié le 10/10/2012 à 18:04

Des scientifiques viennent de montrer qu'un brin d'ADN devient illisible au bout de quelques millions d'années seulement. Un délai bien plus court que les 65 millions d'années qui nous séparent de la disparition des dinosaures.

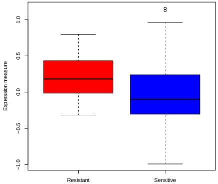
Does it make Jurassic Park unrealistic?

Ancestral genome reconstruction

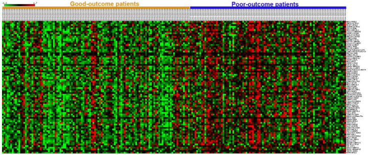


Actually it does. But given enough descendants, we can infer the genome of extinct ancestors (black death, LUCA).

Differential expression testing



Does the expression of one gene change between two groups of samples?



What about the expression of a **set** of genes (*e.g.*, involved in the same biological function)?

Learning with high dimensional data

- Each of these examples involves **complex objects/large numbers of features** for a **restricted number of samples**.
- Intuitively, observing all these characteristics should allow us to predict or understand complex mechanisms.
- We now discuss why this wealth of features can cause trouble in statistical learning.
- Understanding this problem should give more perspective to the tools we will present later.

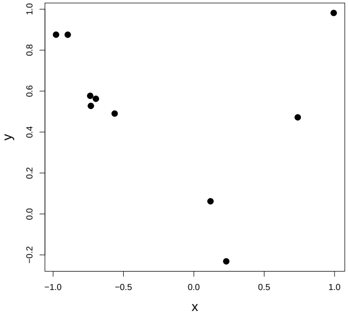
Part II

Overfitting, bias-variance tradeoff: what is the problem?

Short term

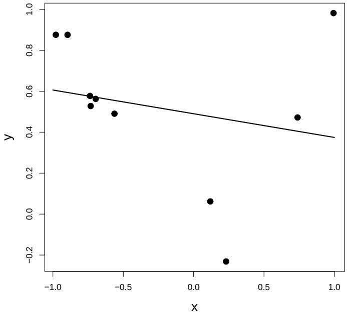
- We start with an informal example.
- We will formalize what we observe later.

Bias-variance tradeoff: intuition



- We observe 10 couples (x_i, y_i).
- We want to estimate y from x.
- Strategy: find f such that $f(x_i)$ is close to y_i.

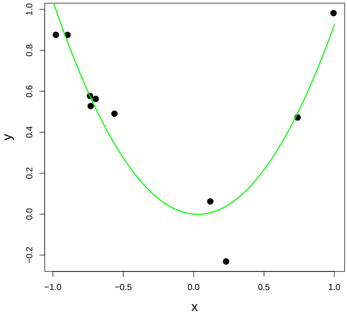
Bias-variance tradeoff: intuition



Find f as a line

$$\min_{f(x)=ax+b} \|Y - f(X)\|^2$$

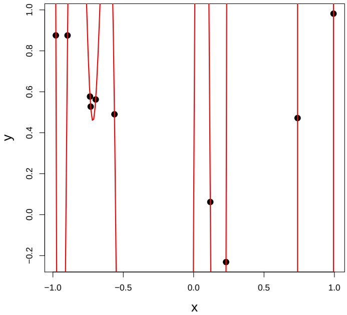
Bias-variance tradeoff: intuition



Find f as a quadratic function

$$\min_{f(x)=ax+bx^2} \|Y - f(X)\|^2$$

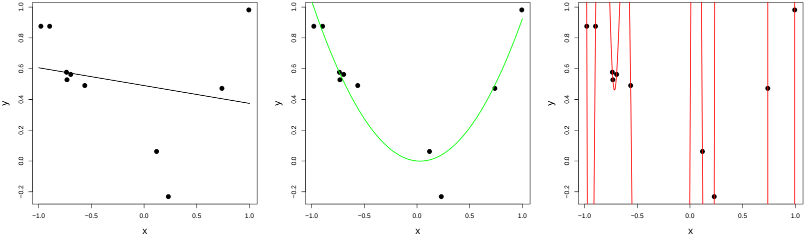
Bias-variance tradeoff: intuition



Find f as a polynomial of degree 10

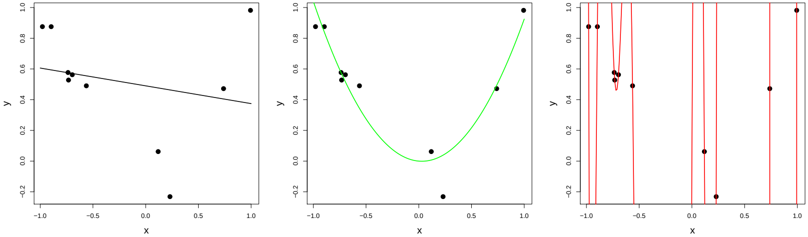
$$\min_{f(x)=\sum_{j=1}^{10} a_j x^j} \|Y - f(X)\|^2$$

Bias-variance tradeoff: intuition



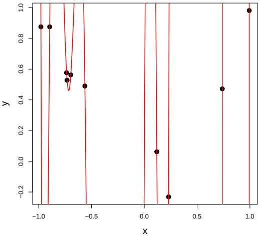
Which function would you trust to predict y corresponding to $x = 0.5$?

Bias-variance tradeoff: intuition



- Reminder: we aim at "finding f such that $f(x_i)$ is close to y_i".
- With the polynomial of degree 10, $f(x_i) - y_i = 0$ for all 10 points.
- There is something wrong with our objective.

Bias-variance tradeoff: intuition

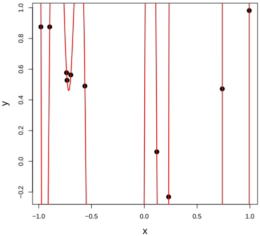


More precisely:

- If we allow any function f, we can find **a lot** of perfect solutions.
- Our actual goal is to estimate y for new points x from the same population :

$$\min_f \mathbb{E}_{(X,Y)} \|Y - f(X)\|^2$$

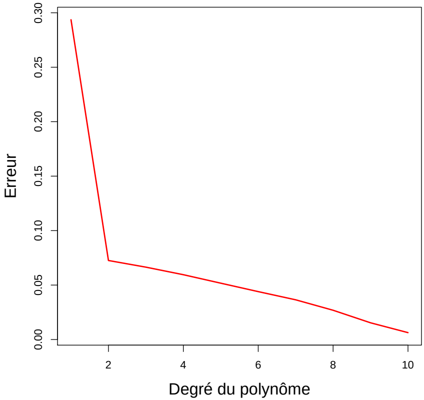
Biais-variance tradeoff: intuition



Even more precisely :

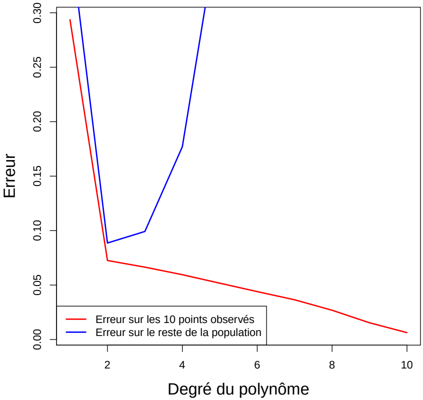
- We did not take into account the fact that our 10 points are a subsample from the population.
- If we sample 10 new points from the same population, the complex functions are likely to change more than the simple ones.
- Consequence: these functions will probably generalize less well to the rest of the population.

Overfitting



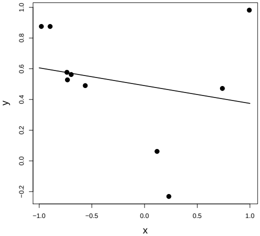
- When the degree increases, the error $\|y - f(x)\|^2$ over the 10 observations always decreases.
- Over the rest of the population, the error decreases, **then increases**.

Overfitting



- When the degree increases, the error $\|y - f(x)\|^2$ over the 10 observations always decreases.
- Over the rest of the population, the error decreases, **then increases**.

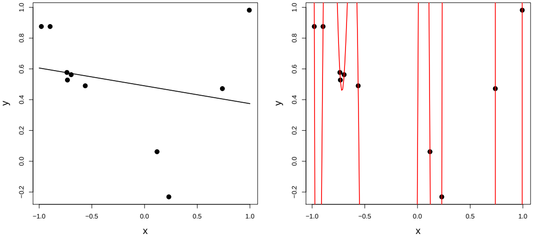
Overfitting



This suggests the existence of a **tradeoff** between two types of errors:

- Sets of functions which are too simple cannot contain functions which explain the data well enough.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample.
- Notion of richness/simplicity will be made very precise later. For now, polynomial degree (number of parameters).

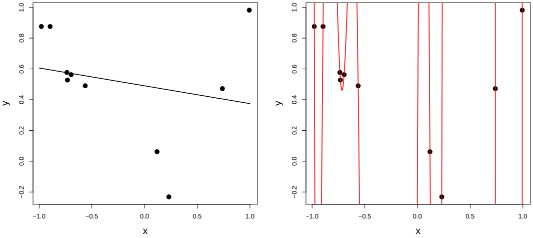
Overfitting



This suggests the existence of a **tradeoff** between two types of errors:

- Sets of functions which are too simple cannot contain functions which explain the data well enough.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample.
- Notion of richness/simplicity will be made very precise later. For now, polynomial degree (number of parameters).

Overfitting



This suggests the existence of a **tradeoff** between two types of errors:

- Sets of functions which are too simple cannot contain functions which explain the data well enough.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample.
- Notion of richness/simplicity will be made very precise later. For now, polynomial degree (number of parameters).

Parenthesis: complexity vs dimension (1/3)

- Our introductive examples had **a large number of descriptors**, and this is a **high dimensional statistics** class.
- This case involves increasingly **complex** functions of a single variable.

Parenthesis : complexity vs dimension (2/3)

- In fact, the two notions are related: here in particular, the three functions are linear in different representations.
- Reminder (linear regression): $\arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2 = (X^\top X)^{-1} X^\top Y$ (if $X^\top X$ is invertible).
- How can we use this fact to compute $\arg\min_{f(x) = \sum_{j=1}^{p} a_i x^j} \|Y - f(X)\|^2$?

Parenthesis : complexity vs dimension (3/3)

- We could have illustrated the same principle using linear functions involving more and more variables.
- Example : predicting a phenotype using the expression of an increasing number of genes.
- We sticked to polynomials, which allow for better visual representations.
- Along this class, the notion of complexity of a set of functions will become more and more precise.
- Complexity is what causes problems for inference, not just dimension.

Second parenthesis : models

- Until now, we did not need to introduce a **model** for the data, *i.e.*, a distribution over $\mathcal{X} \times \mathcal{Y}$:
    - Data could come from any population.
    - The functions we used to predict y can be derived from particular probabilistic models, but this is not necessary (they were in fact historically introduced without a model).
- The objective is not to criticize the use of models, but to show that the tradeoff problem we introduced goes beyond probabilistic models.
- We now show how using a model can give a better insight into the problem.

A little more formally: biais-variance decomposition

- We now assume that the data follow:

$$y = f(x) + \varepsilon, \tag{1}$$

  and $\mathbf{E}[\varepsilon] = 0$.
- Without loss of generality, we consider an estimator $\hat{f}$ of f, function of the data $\mathcal{D} = (x_i, y_i)_{(i=1,\dots,n)}$ generated under (1) (so don't forget: $\hat{f}$ is a random quantity).
- We consider the mean **quadratic error** $\mathbf{E}[(y - \hat{f}(x))^2]$ incurred when using $\hat{f}$ to estimate y from x, generated under (1) but independent from $\mathcal{D}$.
- Expectation is taken over the $(n+1)$ (x, y) pairs : n to build $\hat{f}$, plus the one over which we compute the error.

A little more formally: biais-variance decomposition

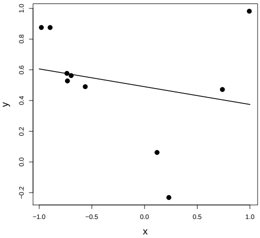
Proposition

Under the previous hypotheses,

$$\mathbf{E}[(y-\hat{f}(x))^2] = \Big(\mathbf{E}[\hat{f}(x)] - f(x)\Big)^2 + \mathbf{E}\left[\Big(\mathbf{E}[\hat{f}(x)] - \hat{f}(x)\Big)^2\right] + \mathbf{E}[(y-f(x))^2]$$

- The first term is the squared bias of $\hat{f}$: the difference between its mean (over the samples in $\mathcal{D}$) and the true f.
- The second term is the variance of $\hat{f}$: how much $\hat{f}$ varies around its average when the data change.
- The third term is the Bayes error, and does not depend on the estimator. The actual quantity of interest is the **excess of risk** $\mathbf{E}[(y-\hat{f}(x))^2] - \mathbf{E}[(y-f(x))^2]$.

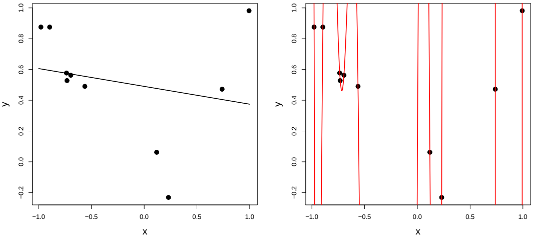
Back to our example



Tradeoff between two types of error:

- Sets of functions which are too simple cannot contain functions which explain the data well enough:
  these sets lead to estimators with a large bias.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample:
  these sets lead to estimators with a large variance.

Back to our example



Tradeoff between two types of error:

- Sets of functions which are too simple cannot contain functions which explain the data well enough:
  these sets lead to estimators with a large bias.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample:
  these sets lead to estimators with a large variance.

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\mathbf{E}[(y - \hat{f}(x))^2] = \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2]$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\begin{aligned}\mathbf{E}[(y - \hat{f}(x))^2] &= \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\ &= \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2]\end{aligned}$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z-\mathbf{E}[Z])^2\right]=\mathbf{E}[Z^2]-\mathbf{E}[Z]^2$

$$\begin{aligned}
\mathbf{E}[(y-\hat{f}(x))^2] =& \mathbf{E}[y^2-2y\hat{f}(x)+\hat{f}(x)^2]\\
=& \mathbf{E}[y^2]-\mathbf{E}[2y\hat{f}(x)]+\mathbf{E}[\hat{f}(x)^2]\\
=& \mathbf{E}[y]^2+\mathbf{E}[(y-\mathbf{E}[y])^2]\\
& -2\mathbf{E}[y]\mathbf{E}[\hat{f}(x)]\\
& +\mathbf{E}[\hat{f}(x)]^2+\mathbf{E}[(\hat{f}(x)-\mathbf{E}[\hat{f}(x)])^2]
\end{aligned}$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z-\mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$
\begin{aligned}
\mathbf{E}[(y-\hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\
=& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2] \\
=& f(x)^2 + \mathbf{E}[(y-f(x))^2] \\
& - 2f(x)\mathbf{E}[\hat{f}(x)] \\
& + \mathbf{E}[\hat{f}(x)]^2 + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2]
\end{aligned}
$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z-\mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\begin{aligned}
\mathbf{E}[(y-\hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\
=& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2] \\
=& f(x)^2 + \mathbf{E}[(y-f(x))^2] \\
& - 2f(x)\mathbf{E}[\hat{f}(x)] \\
& + \mathbf{E}[\hat{f}(x)]^2 + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2] \\
=& \mathbf{E}[(y-f(x))^2] + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2] \\
& + \left(\mathbf{E}[\hat{f}(x)] - f(x)\right)^2
\end{aligned}$$

Biais-variance decomposition : perspective

- Using a (rather general) model, we managed to start formalizing the tradeoff introduced with our example.
- We now generalize this formalization.

A little more generally : structural risk minimization

- We now suppose more generally that the observations are sampled from a joint distribution $\mathbb{P}(x, y)$.
- This does not necessarily mean that we assume a particular probabilistic model: given a deterministic set of couples (x, y), $\mathbb{P}$ can be their empirical distribution.
- We also consider a **loss function**

  $$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

  $L(y, y')$ quantifies the cost of the error made by predicting y' when the true value is y.
- Special case (our example): $L(y, y') = (y - y')^2$.

A little more generally : structural risk minimization

We look for an estimator $f : \mathcal{X} \rightarrow \mathcal{Y}$ minimizing

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x)) d\mathbb{P} = \mathbf{E}[L(y, f(x))]. \tag{2}$$

R is the **risk** of f : the average cost of using f to predict y from x over the joint distribution.

A little more generally : structural risk minimization

- In practice, we cannot compute $R(f)$ because the distribution $\mathbb{P}$ is unknown.
- We therefore use a training set ($\mathcal{D}$ in the previous example) to estimate R, for example through the **empirical risk**:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)). \tag{3}$$

- **Empirical risk minimization** : choose f minimizing $\hat{R}$.
- We saw in our example that minimizing the empirical risk over a "complex/rich" set of functions was not enough to obtain a low risk R.

A little more generally : structural risk minimization

- Assume we minimize the risk over a function space $\mathcal{H}$ (polynomials of a certain degree in our example).
- If R^* is the Bayes risk, we can decompose the **Bayes regret** :

$$R(f) - R^* = \left(R(f) - \inf_{g\in\mathcal{H}} R(g)\right) + \left(\inf_{g\in\mathcal{H}} R(g) - R^*\right). \qquad (4)$$

- The second term is the approximation error: the smallest excess of risk we can reach using a function of $\mathcal{H}$.
- This is a bias term, which does not depend on the data but only on the size of $\mathcal{H}$.
- The first term is the excess of risk of f with respect to the best function in $\mathcal{H}$.

A little more generally : structural risk minimization

- We consider $\hat{f}$ obtained by minimizing the empirical risk over $\mathcal{H}$:

$$\hat{f} \in \arg\min_{g \in \mathcal{H}} \hat{R}(g)$$

- We want to bound the excess of risk $R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \geq 0$
- This term (estimation error) can be decomposed:

$$\begin{aligned} R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \overset{\Delta}{=} & R(\hat{f}) - R(f^*_{\mathcal{H}}) \\ = & R(\hat{f}) - \hat{R}(\hat{f}) \\ & + \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}}) \\ & + \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}). \end{aligned}$$

A little more generally : structural risk minimization

$$\begin{aligned}R(\hat{f}) - \inf_{g\in\mathcal{H}} R(g) =& R(\hat{f}) - R(f^*_{\mathcal{H}})\\ =& R(\hat{f}) - \hat{R}(\hat{f})\\ & + \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}})\\ & + \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}).\end{aligned}$$

- Reminder :
  - R is the **population** risk, $\hat{R}$ the **empirical** risk, an estimator.
  - $\hat{f}$ is the minimizer of $\hat{R}$ over $\mathcal{H}$, $f^*_{\mathcal{H}}$ is the minimizer of R over $\mathcal{H}$.
  - We therefore estimate at two levels: the function f and the risk R.

A little more generally : structural risk minimization

$$
\begin{aligned}
R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) &= R(\hat{f}) - \hat{R}(\hat{f}) \\
&+ \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}}) \\
&+ \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}).
\end{aligned}
$$

- The first term is the difference between the true risk and the estimated risk, for $\hat{f}$.
- This is a complex object to study. **Statistical learning theory** (Vapnik and Chervonenkis) aims at bounding this quantity as a function of n and the complexity of $\mathcal{H}$.
- The second term is nonpositive by construction.
- The third one is easier to control as it involves a deterministic function and the law of large numbers applies.

A little more generally : structural risk minimization

We can however bound the first term:

$$R(\hat{f}) - \hat{R}(\hat{f}) \leq \sup_{f \in \mathcal{H}} \left| \mathbf{E}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right|,$$

and since this quantity also bounds the third term, we get

$$R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \leq 2 \sup_{f \in \mathcal{H}} \left| \mathbf{E}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right|.$$

- This bound of the estimation error suggests that it corresponds to a variance term, which increases with the size of $\mathcal{H}$.
- The more complex $\mathcal{H}$ is, the more likely it is to contain a function for which the empirical risk and the population risk are very different.

A little more generally : structural risk minimization

We can make this notion of size more precise by introducing the **Rademacher complexity** of $\mathcal{H}$:

Definition

Let ϵ_i, $i = 1, \dots, n$ i.i.d such that $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, Z_i, $i = 1, \dots, n$ i.i.d data and $\mathcal{H}$ a space of functions defined over this data, then

$$\mathfrak{R}(\mathcal{H}) = \mathbf{E}_{\epsilon_1^n, Z_1^n}\left[\sup_{f \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(Z_i)\right|\right]$$

is the Rademacher complexity of $\mathcal{H}$.

Intuition: $\mathfrak{R}$ measures the capacity of $\mathcal{H}$ to provide functions which align with noise.

A little more generally : structural risk minimization

We can make this notion of size more precise by introducing the **Rademacher complexity** of $\mathcal{H}$:

Definition

Let ϵ_i, $i = 1, \dots, n$ i.i.d such that $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, Z_i, $i = 1, \dots, n$ i.i.d data and $\mathcal{H}$ a space of functions defined over this data, then

$$\mathfrak{R}(\mathcal{H}) = \mathbf{E}_{\epsilon_1^n, Z_1^n}\left[\sup_{f \in \mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(Z_i)\right|\right]$$

is the Rademacher complexity of $\mathcal{H}$.

This complexity increases with the size of $\mathcal{H}$ and decreases with the size n of the sample.

A little more generally : structural risk minimization

We can bound the mean estimation error in terms of the Rademacher complexity of $\mathcal{H}$.

Proposition

$$\mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y, f(x))] - \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) \right| \leq 2\mathfrak{R}(\mathcal{H}).$$

Therefore,

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}).$$

A little more generally : structural risk minimization

Therefore

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}),$$

- This result illustrates a little more generally the bias variance tradeoff for risk minimization.
- It makes explicit the link between complexity and sample size: lots of points are needed to estimate in large $\mathcal{H}$ (otherwise $\mathfrak{R}(\mathcal{H})$ is large).

ERM consistency and SRM

Therefore

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}),$$

Concretely, this analysis is at the core of two major elements of statistical learning (Vapnik and Chervonenkis, late 60's):

- It is used in learning theory to establish consistency of empirical risk minimization: only families with bounded complexity allow to learn by ERM (are consistent).
- **It also suggests a strategy to design estimators**: build small classes $\mathcal{H}$ which we think contain good approximations.