Intelligent data analysis
From coherent reasoning to universal consistency

Peter Antal antal@mit.bme.hu

Overview

- Prev: Universal theory of induction
  - Best achievable asymptotic(!) performance
- Any finite bound for the performance?
  - No Free lunch
  - From coherent reasoning to repeatibility, independence, concept of true model, and hypothesis testing
  - De Finetti
  - The BIG CHANGE: your/fixed data – all modell vs your/fixed model – all data
- Model complexity
  - Number of free variables
  - VC dimension
- Curse of dimensionality, Bias-variance, Multiple Hypothesis Testing Problem

Universal theory of induction

- Universal distributions

$$m(x) := \sum_{p\,:\,U(p)=x} 2^{-\ell(p)}\,, \qquad -\log m(x) = K(x) + O(1).$$

$$M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}\,, \qquad -\log M(x) = K(x) - O(\log \ell(x))$$

~to predict next symbol find all the compatible Turing machines and weight their prediction with their Kolmogorov complexity...

M. Li and P. M B. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, New York, 2nd edition 1997, and 3rd edition 2008

“NO FREE LUNCH” theory….

The “as if” theorems

de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome. Or more exactly:

1. I.: A coherent system of preferences can be interpreted as probabilities.
2. II.: A coherent system of preferences can be interpreted as utilities (with probabilities).
3. III.: The exchangeability of observations can be interpreted as model averaging.

From ”rational" preferences to probabilities: ”as if" I.

1. Definition. *A decision problem is defined by the elements* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$, *where:*

(i) $\mathcal{E}$ *is an algebra of events,* E_j*;*

(ii) $\mathcal{C}$ *is a set of possible consequences,* c_j*;*

(iii) $\mathcal{A}$ *is a set of possible acts, which are mapping of partitions of the events to consequences;*

(iv) $\leq$ *is a binary preference relation between some of the elements of* $\mathcal{A}$*.*

With further "rational" assumptions on comparability, transitivity, consistency and quantification the following suggestive result can be derived.

1. Proposition. *Given an uncertainty relation* $\leq$*, there exists a unique real number P(E) for each event E (defined as the probability of E) that they are compatible with* $\leq$ *(i.e.* $E \leq F; iff; P(E) \leq P(F)$*) and they form a finitely additive probability measure.*

Consequently, $P(A|\xi)$ denotes the subjective/personal beliefs in a given space-time-information context ξ vs. the "frequentist" interpretation that $P(A) \triangleq \lim_{N\to\infty} N_A/N$.

From preferences to utilities: ”as if" II.

The parallel result for the existence and uniqueness of utilities (or losses) is stated only for decision problems with bounded consequences.

2. Proposition. *For any decision problem* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$ *with bounded consequences* $c_* < c^*$,

(i) for all c, $u(c|c_*, c^*)$ *exists and unique;*

(ii) the value of $u(c|c_*, c^*)$ *is unaffected by the assumed occurrence of an event G;*

(iii) $0 = u(c_*|c_*, c^*) \leq u(c|c_*, c^*) \leq u(c^*|c_*, c^*) = 1.$

From exchangeability to parameters and ”i.i.d." : ”as if" III.

3. Proposition. *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities with probability measure P, that is for any n and permutation $\pi(1), \ldots, \pi(n)$ the joint mass function of P $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$, there exists a distribution function $\mathcal{Q}$ such that $p(x_1, \ldots, x_n)$ has the form*

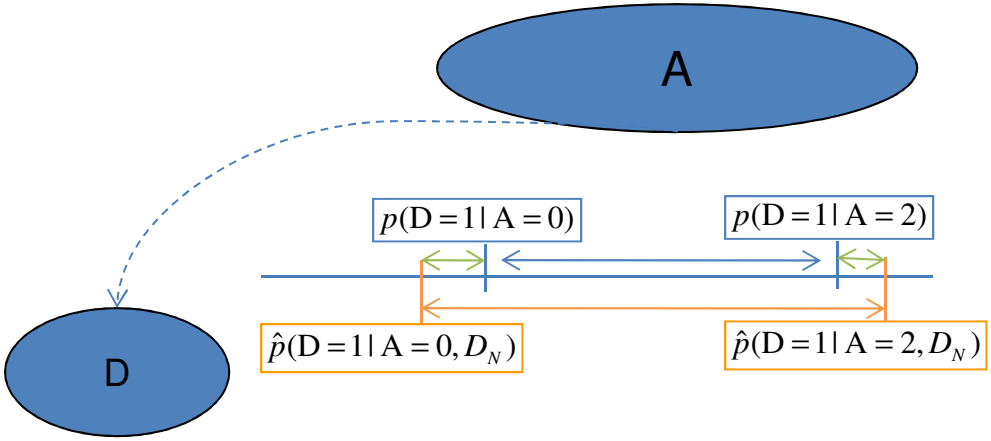
$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta^{1-x_i}) d\mathcal{Q}(\theta),$$

where,

$$\mathcal{Q}(\theta) = \lim_{n\to\infty} P[y_n/n \leq \theta],$$

with $y_n = x_1 + \cdots + x_n$, and $\theta = \lim_{n\to\infty} y_n/n$.

Fundamental questions in statistics



Real difference

Estimation errors

Estimated difference

Relative frequencies: $\hat{p}(\mathrm{D}=1|\mathrm{A}=0,D_N)=\hat{p}(\mathrm{D}=1,\mathrm{A}=0,D_N)/\hat{p}(\mathrm{A}=0,D_N)=N_{D=1,A=0}/N_{A=0}$

Law of large numbers: $p(\mathrm{D}=1|\mathrm{A}=0)\approx N_{D=1,A=0}/N_{A=0}$

Estimation error because of finite data D_N: $\hat{p}(\mathrm{D}=1|\mathrm{A}=0,D_N)$ - $p(\mathrm{D}=1|\mathrm{A}=0)$

Central limit th. (asymptotic), Inequalities for finite(!) data (ε accuracy,δ confidence)

sample complexity: $N_{\varepsilon,\delta}$ $p(D_{N_{\varepsilon,\delta}}:\varepsilon<|\hat{p}(\mathrm{D}=1|\mathrm{A},D_{N\varepsilon,\delta})-p(\mathrm{D}=1|\mathrm{A})|)<\delta$

Miért működik a tanulás: számítási tanulási elmélet

Tanulás: a tapasztalatok eredményeképp javítjuk viselkedésünket.

De hogyan tudhatja valaki, hogy a tanulási algoritmusa olyan elméletet eredményezett, amely helyesen fogja megjósolni a jövőt?

Az induktív tanulás fogalmaival:

honnan tudjuk, hogy a *h* hipotézis jól közelíti az *f* célfüggvényt, ha nem ismerjük *f*-et?

számítási tanulás elmélet

Az alapelv – bármely, súlyos hibákkal terhelt hipotézis már kis számú példa vizsgálata után is szinte bizonyosan „megbukik", mivel nagy valószínűséggel legalább egy helytelen eredményt fog jósolni.

Így valószínűtlen, hogy súlyosan hibás legyen bármely olyan hipotézis, amely egy kielégítően nagy tanuló példahalmazzal konzisztens,

Valószínűleg Közelítőleg Helyes
(Probably Approximatley Correct).
VKH-tanulás (PAC-learning)

De hány példára van szükség a biztonsághoz?

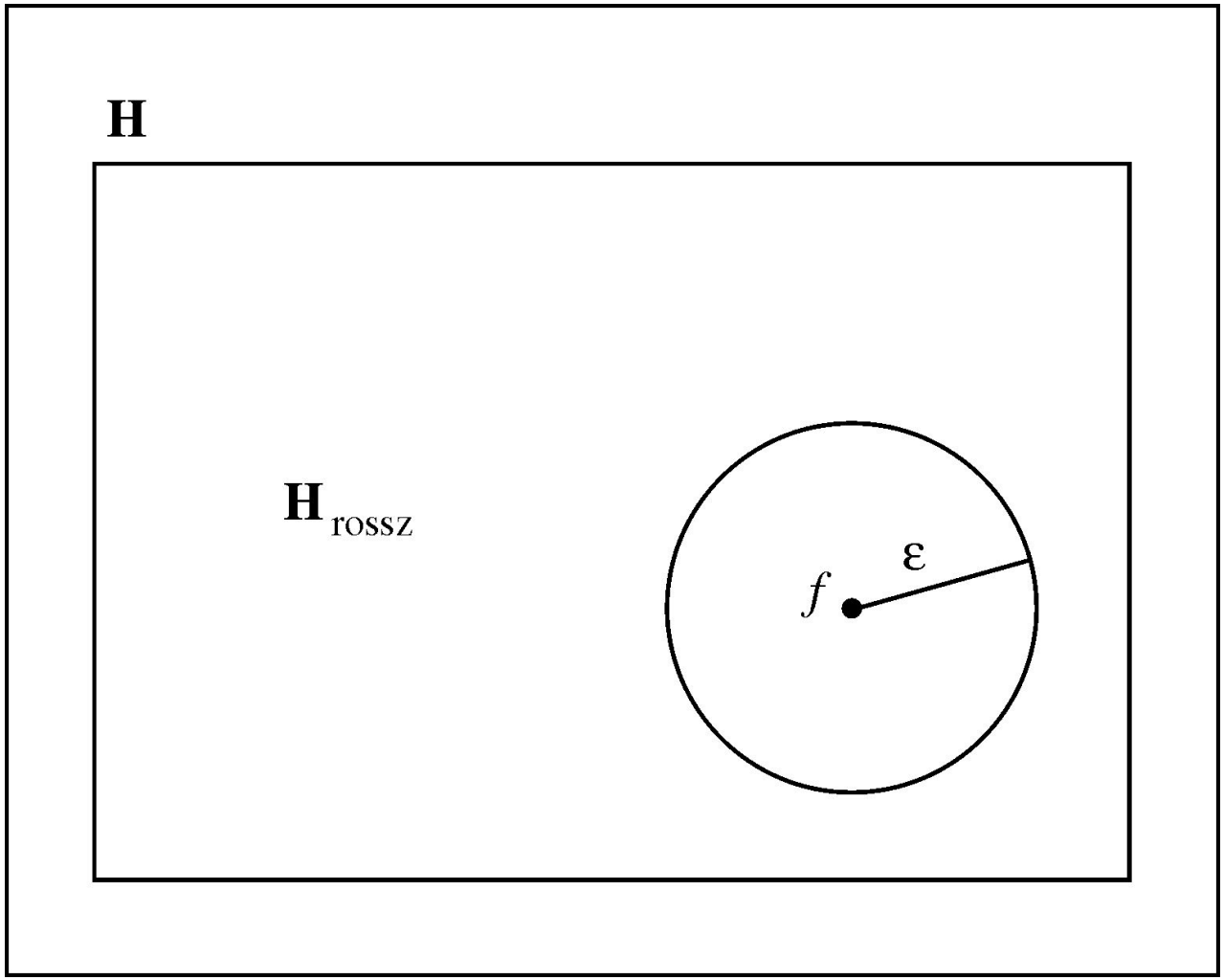
Hány példára van szükség?

X az összes lehetséges példák halmaza.

D a példák eloszlása.

H az összes lehetséges hipotézisek halmaza.

m a tanuló halmaz példáinak száma.



Tegyük fel, hogy a keresett, valódi *f* függvény eleme **H**-nak.

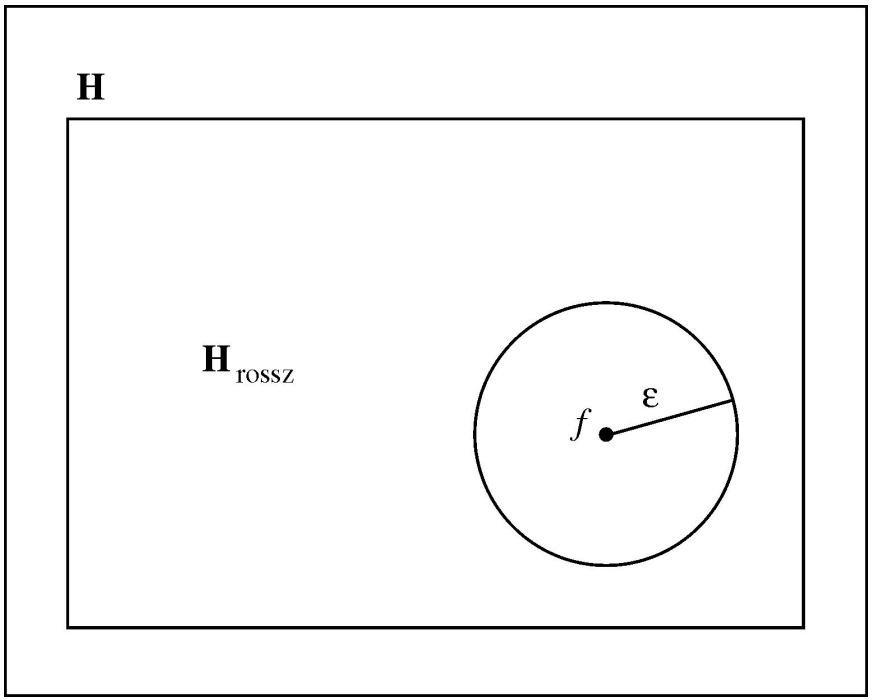
h hipotézis *f* függvényhez képesti **hibája** a *D* eloszlású példahalmazon:

$$\text{hiba}(h) = P(h(x) \neq f(x) \mid x \text{ a } D \text{ eloszlású példáhalmaz eleme})$$

h hipotézist **közelítőleg helyes**, ha $\text{hiba}(h) \leq \varepsilon$, ahol ε egy kis értékű konstans.

***m* példa vizsgálata után nagy valószínűséggel az összes konzisztens hipotézis közelítőleg helyes lesz.**

A közelítőleg helyes hipotézis - „közel" van a hipotézis térben a valós függvényhez.



hipotézis **H** halmazát két részre bontjuk:

az *f* körül felvett ε**-gömb**, és a maradék, $\mathbf{H}_{\mathbf{rossz}}$.

Valószínűség:

egy „alapvetően rossz" *hr* hipotézis az első *m* példával konzisztens legyen.

Tudjuk, hogy $\text{hiba}(h_r) > \varepsilon$, így annak valószínűsége, hogy bármelyik adott példára jó eredményt ad $\leq (1-\varepsilon)$.

Az *m* példára vonatkozó korlát:

$$P(h_r \text{ jó eredményt ad } m \text{ példára}) \leq (1-\varepsilon)^m$$

Ahhoz, hogy a $\mathbf{H}_{\mathbf{rossz}}$ térrészben legyen konzisztens hipotézis, legalább egy a $\mathbf{H}_{\mathbf{rossz}}$ hipotézisei közül konzisztens kell legyen.

Ennek előfordulási valószínűsége:

$$P(\mathbf{H}_{\mathbf{rossz}} \text{ tartalmaz konzisztens hipotézist}) \leq \left|\mathbf{H}_{\mathbf{rossz}}\right|(1-\varepsilon)^m \leq \left|\mathbf{H}\right|(1-\varepsilon)^m$$

Legyen ez a valószínűség valamilyen kis értékű δ konstans:

$$|\mathbf{H}|(1-\varepsilon)^m \leq \delta$$

Ezt akkor érhetjük el, ha az algoritmusnak

$$m \geq \frac{1}{\varepsilon}(\ln\frac{1}{\delta} + \ln|\mathbf{H}|)$$

példát mutatunk.

Ha egy tanuló algoritmus olyan hipotézist ad, amely ennyi példa esetén konzisztens, akkor ennek a hipotézisnek legalább $1-\delta$ valószínűséggel a hibája legfeljebb ε.

Más szavakkal valószínűleg közelítőleg helyes.
E megkívánt példaszám, amely ε és δ függvénye,
a hipotézis tér **minta komplexitása**.

Egy **függvény megtanulható** példákból adott ε és δ szinten, ha a szükséges példaszám probléma paramétereinek (ε, δ, **H**) polinomiális függvénye,

azaz, ha (általában) a hipotézis tér **minta komplexitása exponenciálisnál kisebb**.

Kulcskérdés a hipotézis tér mérete.

Ha **H** az *n* attribútum esetén felvehető Boole függvények halmaza, akkor $|\mathbf{H}| = 2^{2^n}$

Így a hipotézis tér minta komplexitása 2^n szerint nő.

Legyen $\varepsilon = \delta = 10^{-4}$, és n = 2, $m \geq 5000$

n = 10 $m \geq 3 \times 10^6$

Sok példa: **belső** akadályok = **erőforrások**

külső akadályok = **környezet** dinamizmusa

$$|\mathbf{H}| = 2^{2^n}$$

Az előttünk álló dilemma a következő:

ha nem korlátozzuk a tanuló algoritmus által hipotézisként figyelembe vehető függvények terét, akkor az algoritmus nem lesz képes tanulni,

ha viszont korlátozzuk, akkor lehet, hogy éppen a valódi, keresett függvényt zárjuk ki.

Két út van, hogy „kimeneküljünk" ebből a csapdából.

1. ha ragaszkodunk ahhoz, hogy az algoritmus ne csupán valamilyen konzisztens hipotézist adjon meg, hanem lehetőleg a **legegyszerűbbet**, de legtöbb esetben a legegyszerűbb hipotézis megtalálásának problémája **kezelhetetlen**.

2. **legtöbb esetben** nincs szükségünk a Boole függvények teljes kifejező erejére, ennél **kötöttebb** nyelvekkel is megoldhatók feladataink.

(heurisztikus hozzáállás)

Döntési listák tanulása

A **döntési lista** egy kötött forma logikai kifejezése.

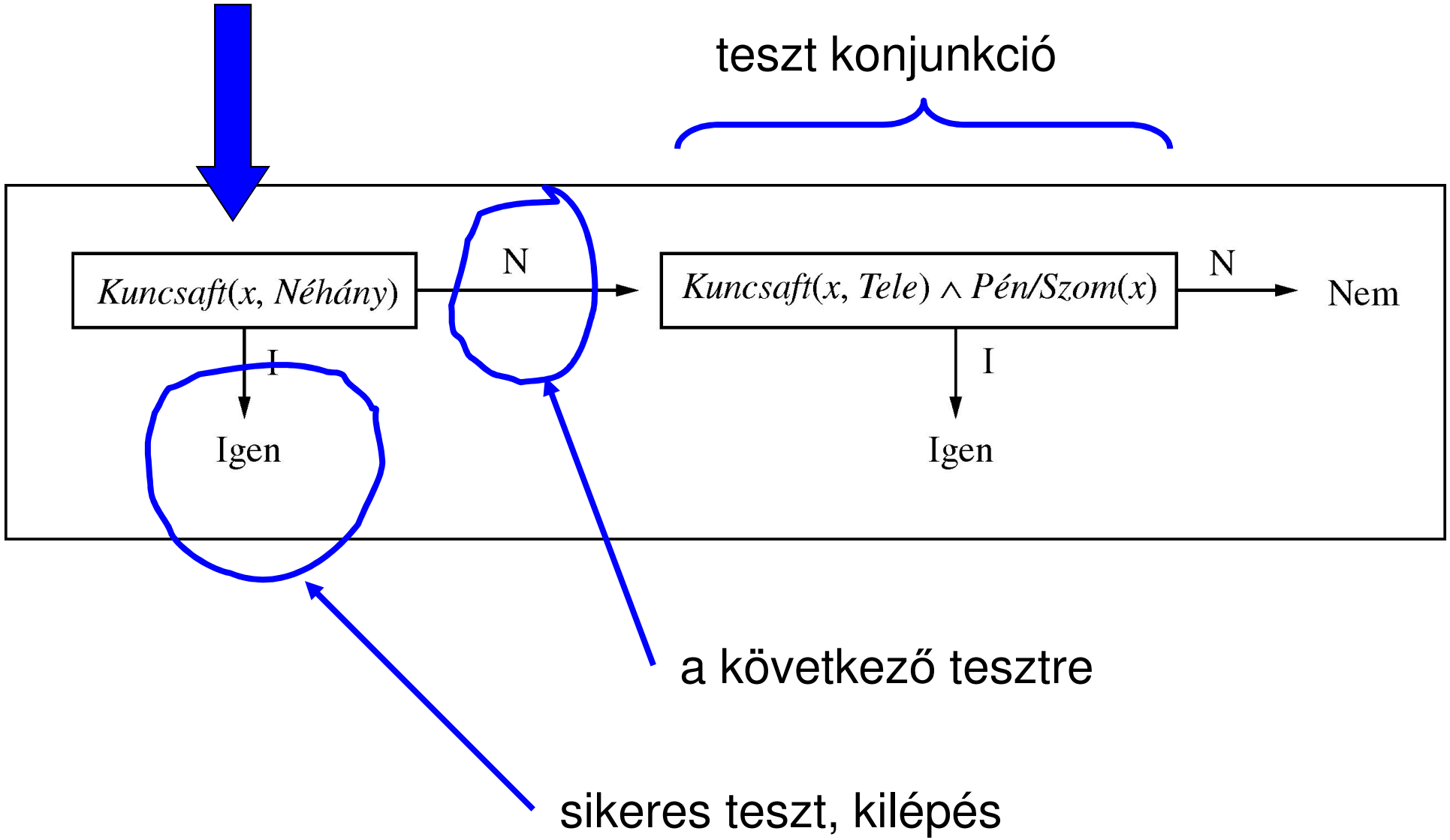
Tesztek sorozata: tesztek mindegyike literálok **konjunkciója**.

Ha egy tesztet sikeresen végrehajtunk **egy példa leírásán**, akkor a döntési lista specifikálja az eredmény értékét.

Ha a teszt sikertelen, akkor a feldolgozás a lista **következő** tesztjével folytatódik.

A döntési lista a döntési fára emlékeztet,
de **globális struktúrája egyszerűbb**,
míg az **egyes tesztek bonyolultabbak**.

Példa beléptetése



Döntési listák tanulása

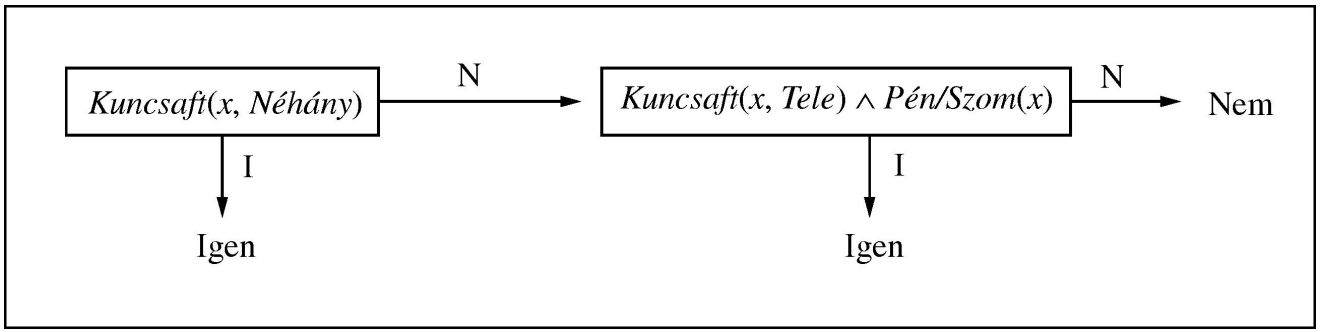
Ha tetszőleges méretű teszteket megengedünk, akkor a döntési lista bármely Boole függvény reprezentálására képes.

(és ott vagyunk, ahol voltunk)

Ha az egyes teszteket legfeljebb ***k*** literálban korlátozzuk, akkor a tanuló algoritmus kisszámú példa alapján is képes lesz sikeresen általánosítani.

Az ilyen nyelvet ***k*-DL** nyelvnek nevezzük.

Ábra:
2-DL példa



Döntési listák tanulása

***k*-DL** $\supset$ ***k*-DT** nyelvek halmaza
(az összes, legfeljebb k mélységű döntési fával leírható nyelvek halmaza)

egy partikuláris k-DL nyelv függ attól is, hogy milyen attribútumokkal írtuk le a példákat

az n Boole attribútumot használó nyelv: ***k*-DL(*n*)**

Megtanulható?

Döntési listák tanulása

a nyelvbe tartozó hipotézisek száma:

a tesztek nyelve – *n* attribútumon felvett legfeljebb *k* literális konjunkciója – *Conj*(*n,k*).

minden egyes teszt *Igen* vagy *Nem* értékkel lehet jelen, vagy pedig hiányozhat a döntési listáról, így legfeljebb

$$3^{|Conj(n,k)|}$$

eltérő komponens teszt halmaz van.

Döntési listák tanulása

Ezen teszt halmazok bármelyike tetszőleges sorrendben tartalmazhatja a teszteket, tehát:

$$\left|k - \mathrm{DL}(n)\right| \leq 3^{\left|Conj(n,k)\right|}\left|Conj(n,k)\right|!$$

Az *n* attribútumból képzett *k* literál lehetséges konjunkcióinak száma:

$$\left|Conj(n,k)\right| = \sum_{i=0}^{k}\binom{2n}{i} = O(n^k)$$

Döntési listák tanulása

Így némi átalakítás után:

$$\left|k - \mathrm{DL}(n)\right| = 2^{O(n^k \log_2(n^k))}$$

egy *k*-DI nyelv VKH-tanulásához szükséges példák száma *n* polinomja:

$$m \geq \frac{1}{\varepsilon}(\ln\frac{1}{\delta} + O(n^k \log_2(n^k)))$$

Learnability, PAC, VC

- Bernstein,..Hoeffding inequality
- Cover
- Stone
- Vapnik-Chervonenkis dimension
- Valiant
- A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth
- ...

Concepts in concept learning

- Probability of error for a classifier g: L(g)=p(g(X)≠Y)
- The best possible classifier: $g^* = \underset{g:\mathcal{R}^d \to \{1,\ldots,M\}}{\arg\min} \mathbf{P}\{g(X) \neq Y\}$
- The best achievable error is the Bayes error: L*=L(g*)
- Classifier vs. rule:

A classifier is constructed on the basis of $X_1, Y_1, \ldots, X_n, Y_n$ and is denoted by g_n; Y is guessed by $g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$. The process of constructing g_n is called *learning*, *supervised learning*, or *learning with a teacher*. The performance of g_n is measured by the conditional *probability of error*

$$L_n = L(g_n) = \mathbf{P}\{g_n(X; X_1, Y_1, \ldots, X_n, Y_n) \neq Y | X_1, Y_1, \ldots, X_n, Y_n)\} .$$

An individual mapping $g_n : \mathcal{R}^d \times \{\mathcal{R}^d \times \{1, \ldots, M\}\}^n \longrightarrow \{1, \ldots, M\}$ is still called a *classifier*. A sequence $\{g_n, n \geq 1\}$ is called a *(discrimination) rule*. Thus, classifiers are functions, and rules are sequences of functions.

Universal consistency

If we are given a sequence $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of training data, the best we can expect from a classification function is to achieve the Bayes error probability L^*. Generally, we cannot hope to obtain a function that exactly achieves the Bayes error probability, but it is possible to construct a sequence of classification functions $\{g_n\}$, that is, *a classification rule*, such that the error probability

$$L_n = L(g_n) = \mathbf{P}\{g_n(X, D_n) \neq Y | D_n\} \qquad \lim_{n \to \infty} \mathbf{P}\{L_n - L^* > \epsilon\} = 0.$$

gets arbitrarily close to L^* with large probability (that is, for "most" D_n). This idea is formulated in the definitions of *consistency*:

DEFINITION 6.1. (WEAK AND STRONG CONSISTENCY). *A classification rule is consistent (or asymptotically Bayes-risk efficient) for a certain distribution of* (X, Y) *if*

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X, D_n) \neq Y\} \to L^* \quad \text{as } n \to \infty,$$

and strongly consistent if

$$\lim_{n \to \infty} L_n = L^* \quad \text{with probability 1.}$$

DEFINITION 6.2. (UNIVERSAL CONSISTENCY). *A sequence of decision rules is called universally (strongly) consistent if it is (strongly) consistent for any distribution of the pair* (X, Y).

Learnability

- 1-NN: $\limsup_{n \to \infty} \mathbf{E} L_n \leq 2L^*$
- For a function class C: $\mathbf{P}\{L_n > L + \epsilon\} \leq 8\left(n^V + 1\right) e^{-n\epsilon^2/128}$
- where $L \stackrel{\text{def}}{=} \inf_{g_n \in C} \mathbf{P}\{g_n(X) \neq Y\}$
- Empirical error and its estimation:

$$\mathbf{P}\{|\widehat{L}_n - L_n| > \epsilon\} \leq \frac{6k+1}{n\epsilon^2}$$

- Universal consistency

$$\mathbf{P}\{L_n - L^* > \epsilon\} \leq e^{-cn\epsilon^2}, \; n \geq N(\epsilon)$$

Non-learnability

- No universal rate of convergence:

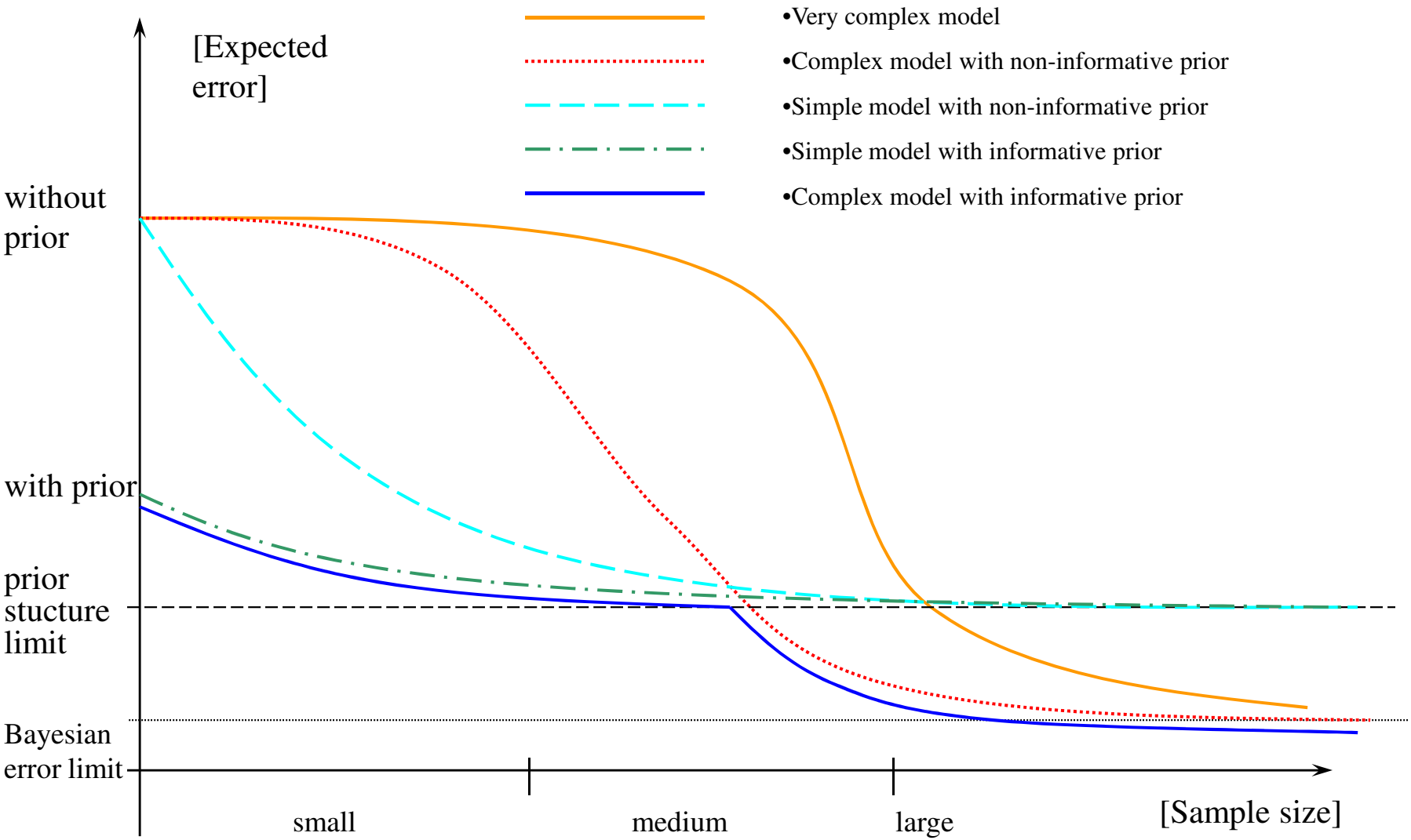
$$\liminf_{n\to\infty} \sup_{\text{all distributions of } (X,Y) \text{ with } L^*+\epsilon<1/2} \mathbf{P}\{L_n \geq L^* + \epsilon\} > 0$$

- L* cannot be estimated universally well.
- No best classifier (E[L(g_n)] cannot be dominant wrt p(X,Y))

- L. Devroye and L. Gy{\"o}rfi and G. Lugosi: A Probabilistic Theory of Pattern Recognition“, 1996

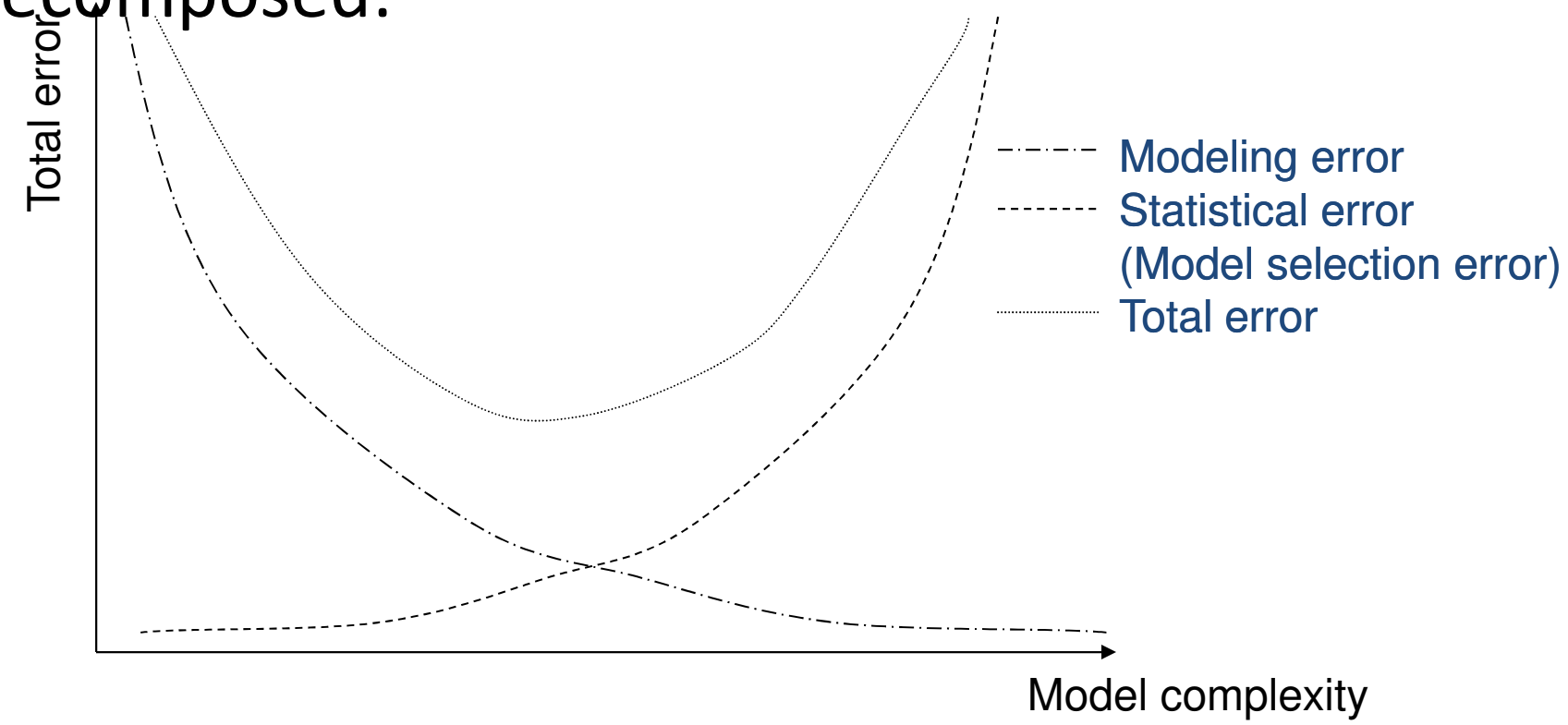
All chapters, without exception, are unashamedly theoretical. We did not scar the pages with backbreaking simulations or quick-and-dirty engineering solutions.

Learning characteristics of various methods



The bias-variance dilemma

- For a given sample size the error is decomposed:



Frequentist vs Bayesian statistics

Frequentist	Bayesian
Considering all data	YOUR DATA
CONSIDERING YOUR HYPOTHESIS	Averaging over models
-	Prior probabilities
Null hypothesis	-
Indirect: proving by refutation	Direct
Model selection	Model averaging
Likelihood ratio test	Bayes factor
p-value	-!
-!	Posterior probabilities
Confidence interval	Credible region
Significance level	Optimal decision based on Exp.Util.
Multiple testing problem	Remains, so ➔ **complex model**
Model complexity dilemma	**Best achievable alternative**

The hypothesis testing framework

- Terminology:
  - False/true x positive/negative
  - Null hypothesis: independence

reported	Ref.:0/N	Ref.1/P
0/N	TN	FN
1/P	FP	TP

  - Type I error/error of the first kind/α error/FP: $p(\neg H_0 | \underline{H_0})$
    - Specificity: $p(H_0 | \underline{H_0}) = 1-\alpha$
    - Significance: α
    - p-value: „probability of more extreme observations in repeated experiments”
  - Type II error/error of the second kind/β error/FN: $p(H_0 | \neg \underline{H_0})$:
    - Power or sensitivity: $p(\neg H_0 | \neg \underline{H_0}) = 1-\beta$

reported	Ref. $\underline{H_0}$	Ref.:$\neg \underline{H_0}$
H_0		Type II
$\neg H_0$	Type I („false rejection”)	

Multiple testing problem (MTP)

- If we perform N tests and our goal is
  - $p(\text{FalseRejection}_1 \text{ or } \ldots \text{ or } \text{FalseRejection}_N) < \alpha$
- then we have to ensure, e.g. that
  - for all $p(\text{FalseRejection}_i) < \alpha/N$

➔loss of power!

E.g. in a GWA study N=100,000, so huge amount of data is necessary….(but high-dimensional data is only relatively cheap!)

Corrections for multiple testing

I have 1,000,000 hypotheses that are not mutually exclusive.

1. I test them all.
   Correction?

2. I plan to test them all, but I run out of resources after testing only one of them.
   Correction?

3. I test one of them, and a year later test the others.
   Correction? If so, when?

4. I only test the first one because that is the one I suspect.
   Correction?

5. I run an algorithm that prunes unlikely hypotheses, keeping only 100,000.
   Correction for 100,000 or for 1,000,000 hypotheses?

(R.Neopolitan, 2010)