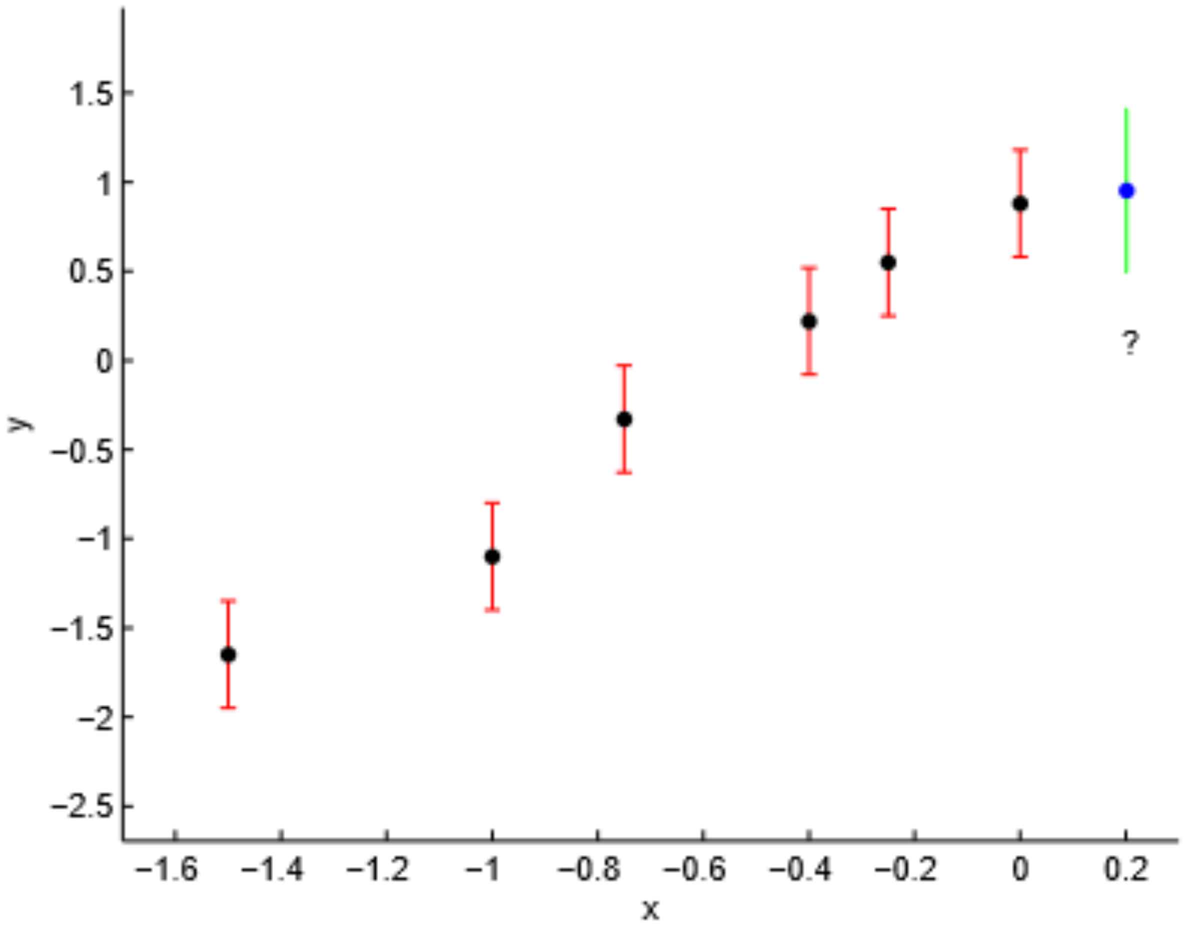
Instituto de
Matemática

Modelagem preditiva, Navalha de Occam e Processos Gaussianos para Machine Learning.

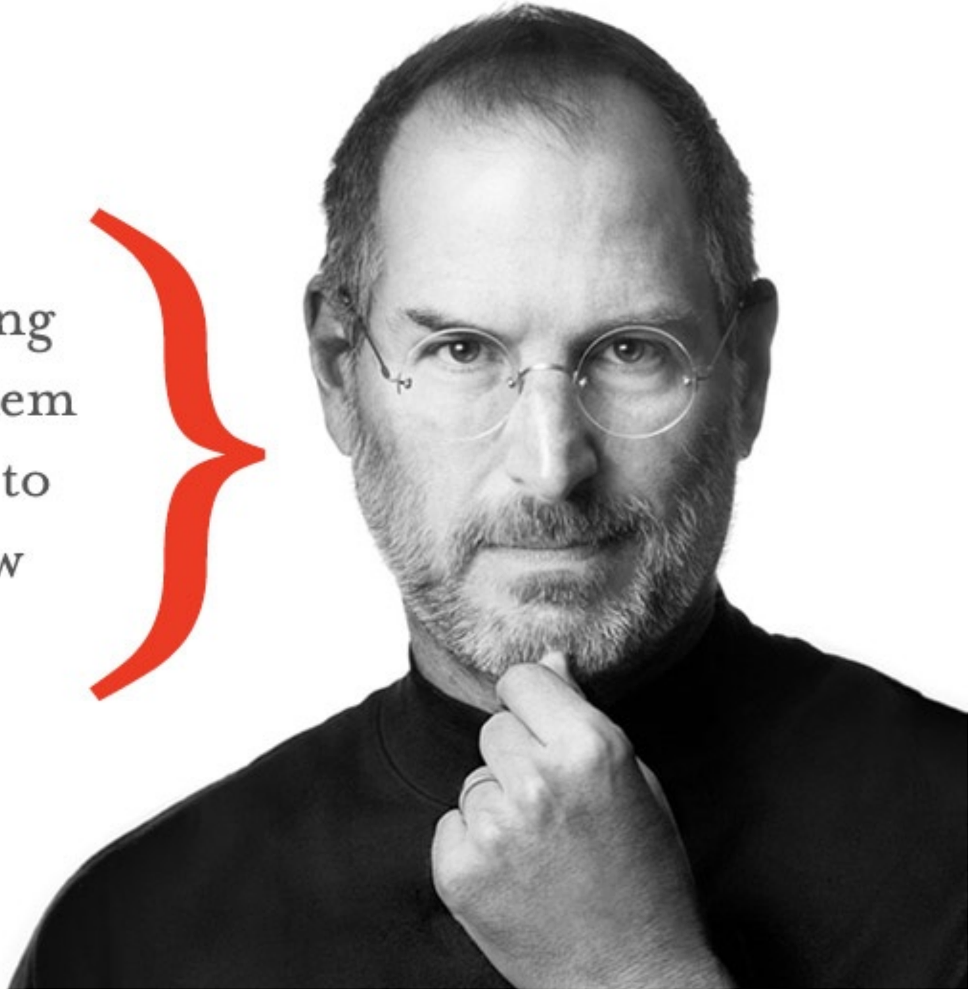
Fabio Ramos
Departamento de Matemática Aplicada
Instituto de Matemática
UFRJ

O OBJETIVO

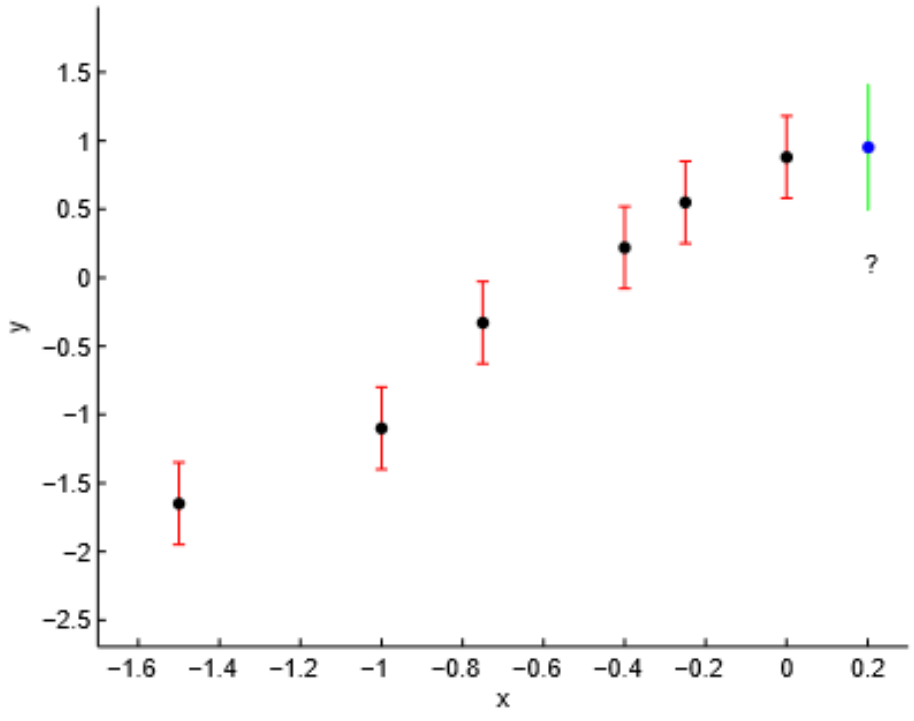


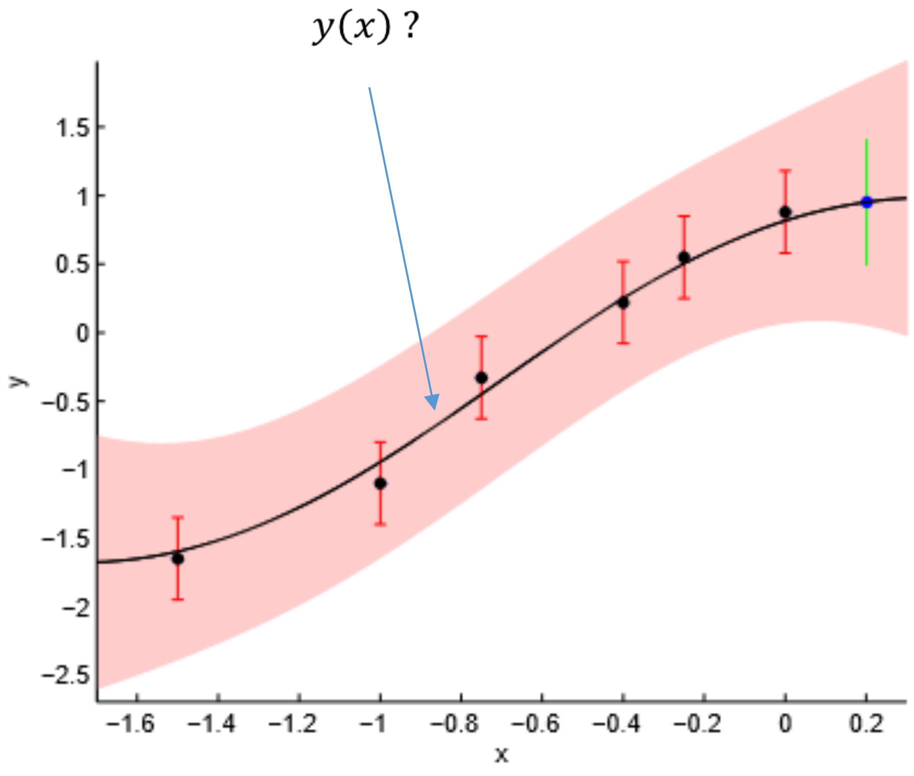
You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future.

–Steve Jobs

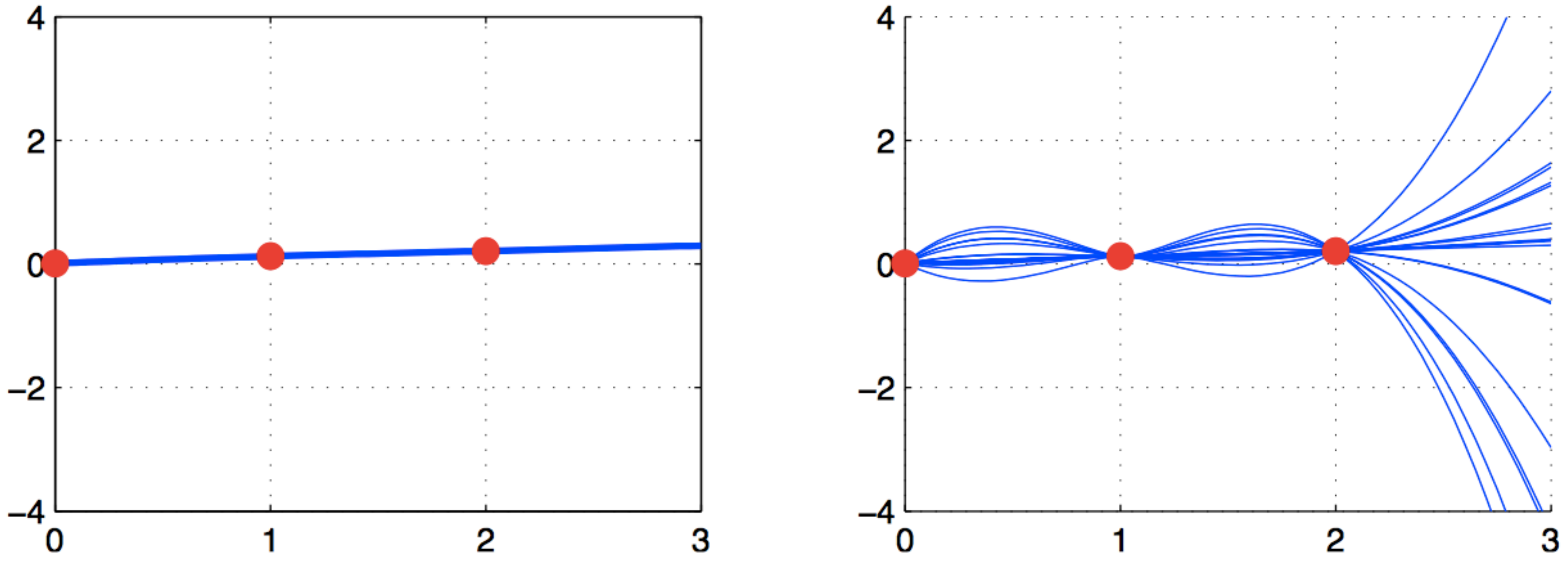


O OBJETIVO

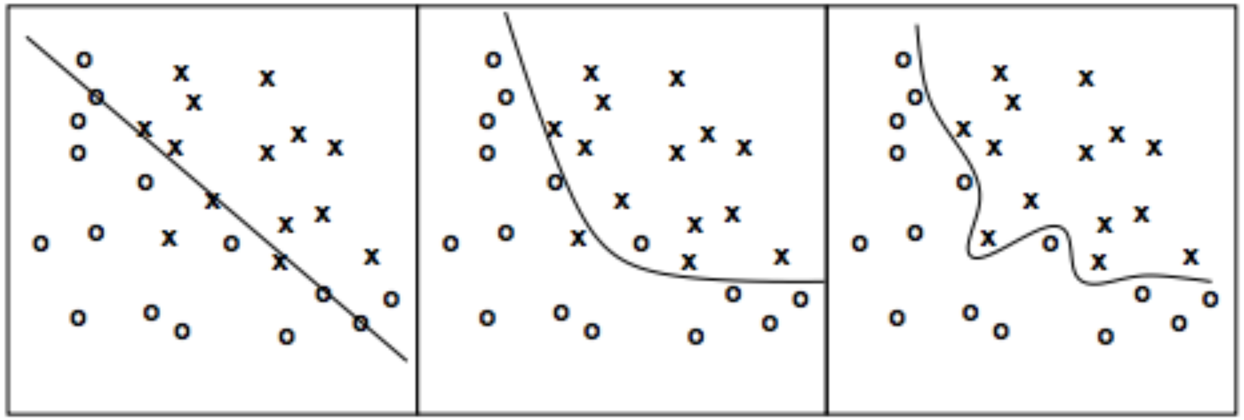




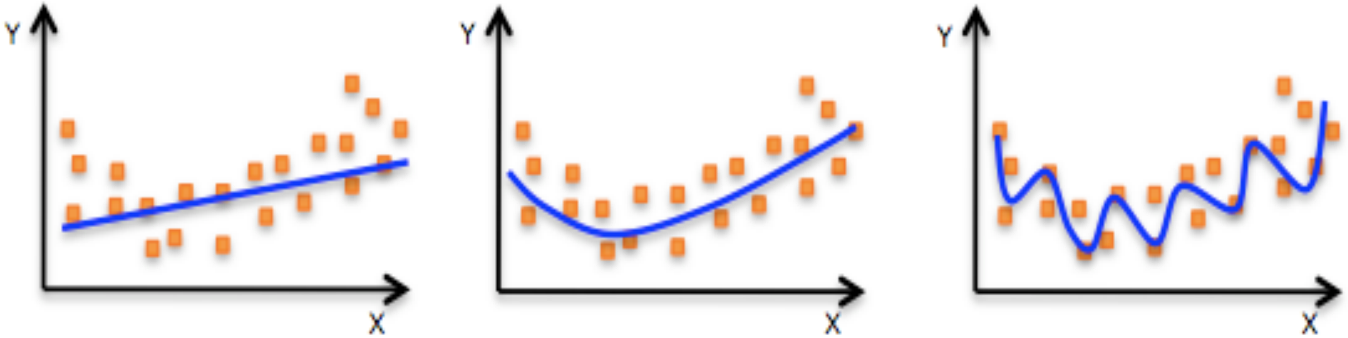
Em qual modelo acreditar ?



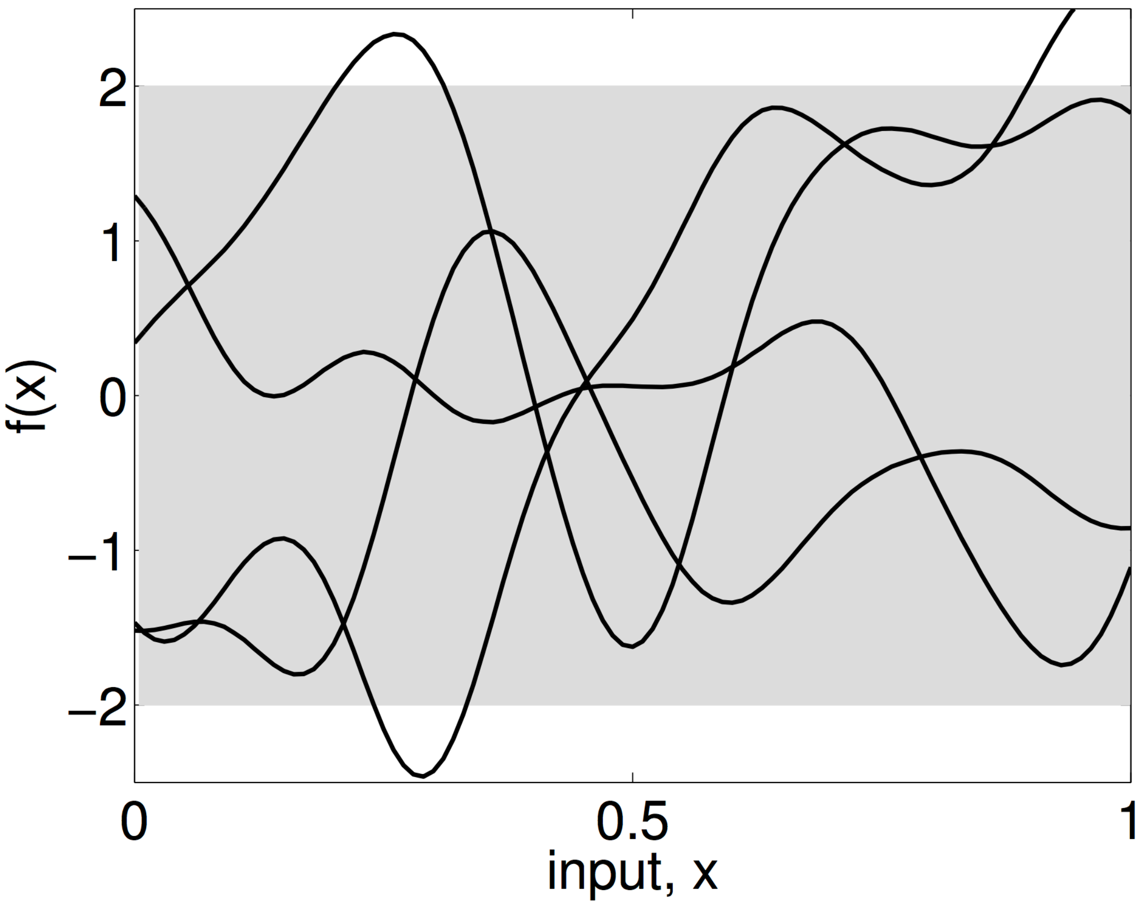
Navalha de Occam - A hipótese mais provável é a mais simples e consistente com os dados



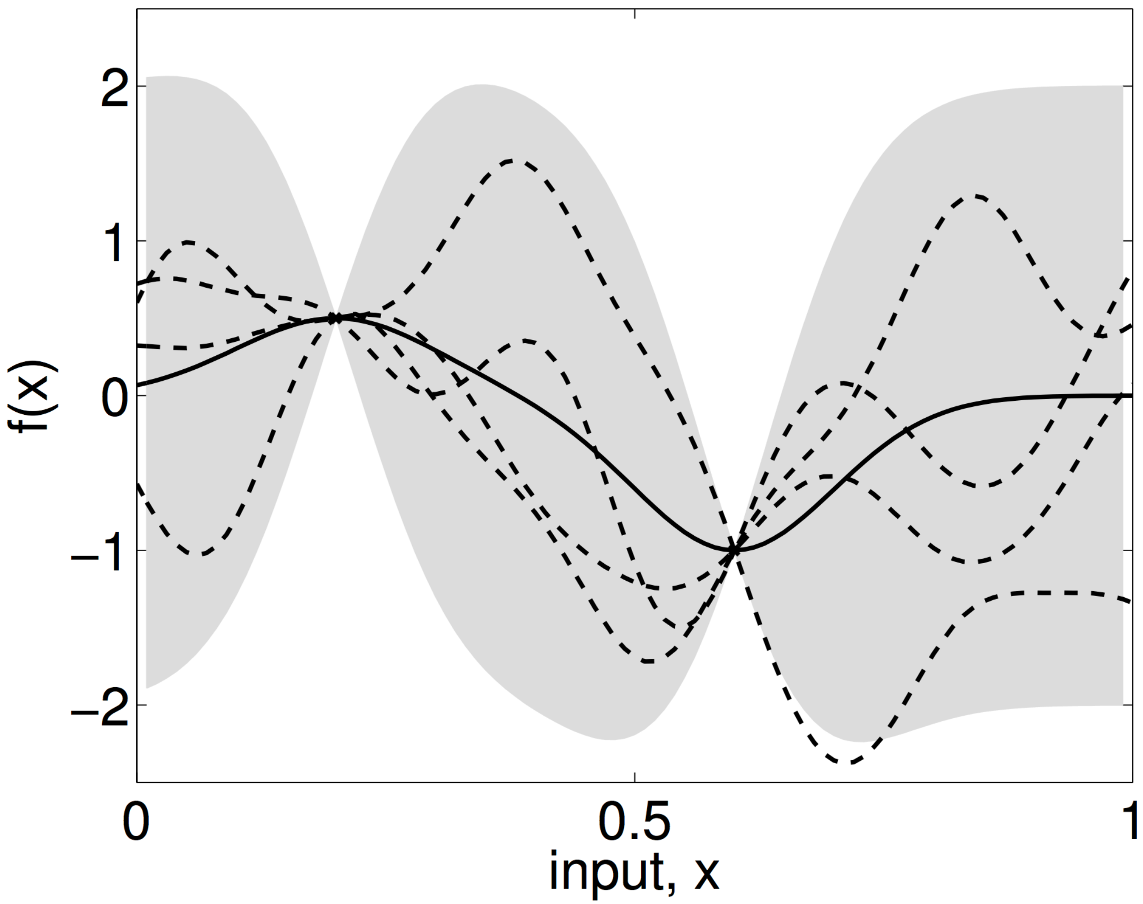
Inadequado | Solução de Compromisso | Overfitting



Gaussian Processes - Priors over functions



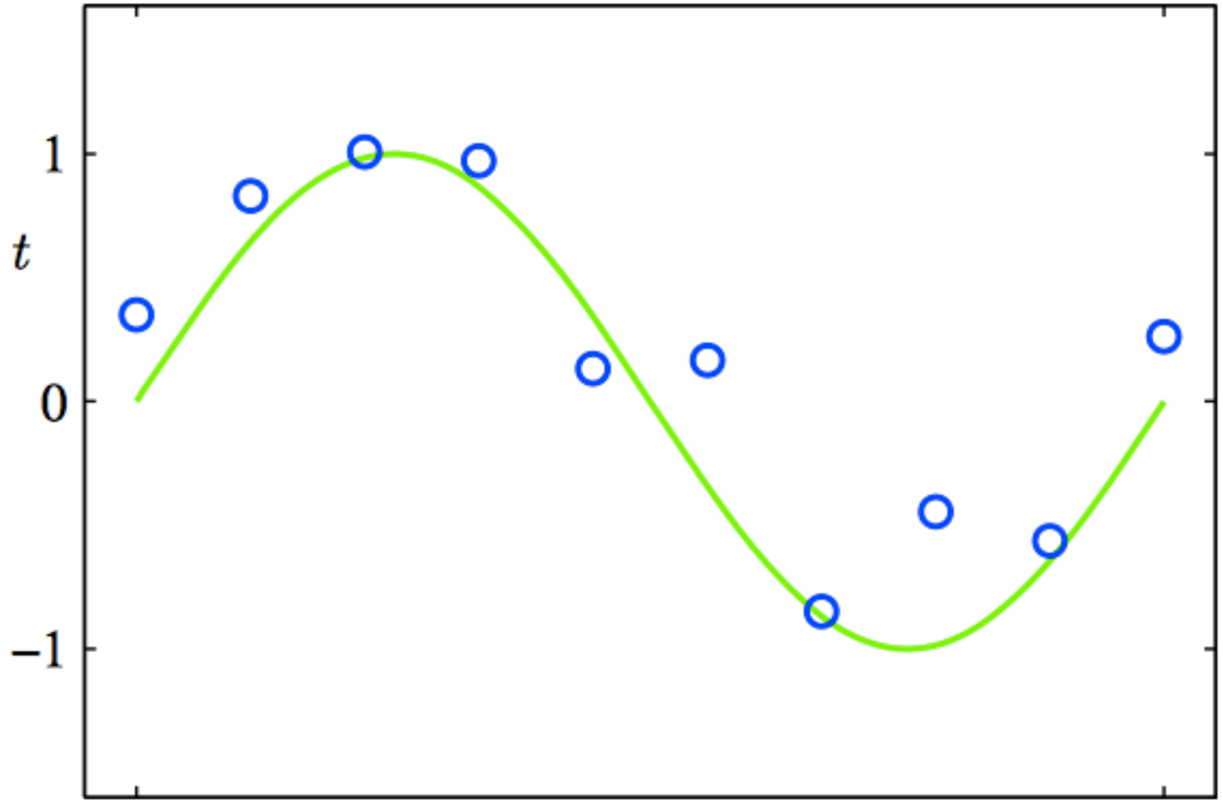
(a), prior



(b), posterior

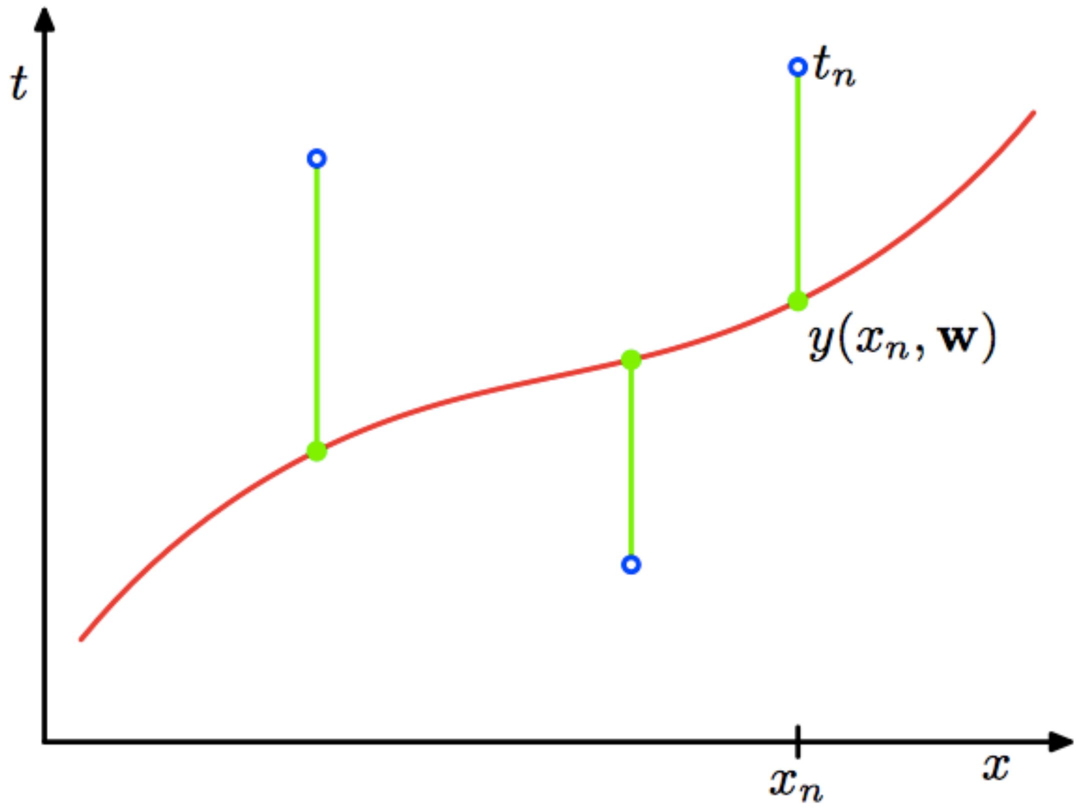
Interpolação polinomial

SINAL (senoidal) + RUÍDO



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

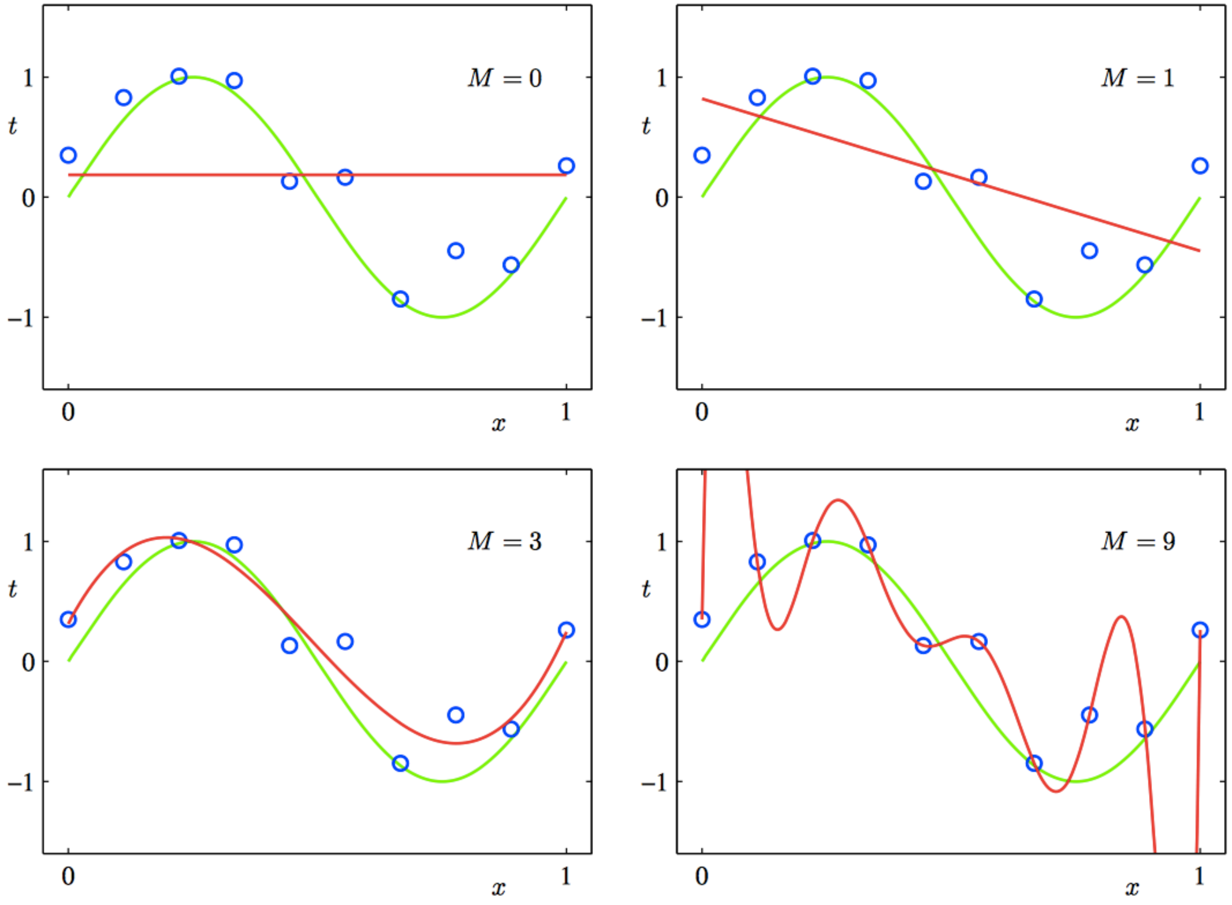
Ajustando: erro médio quadrático



$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \longrightarrow E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

minimizador $\mathbf{w}^*$

Melhor Interpolação Polinomial

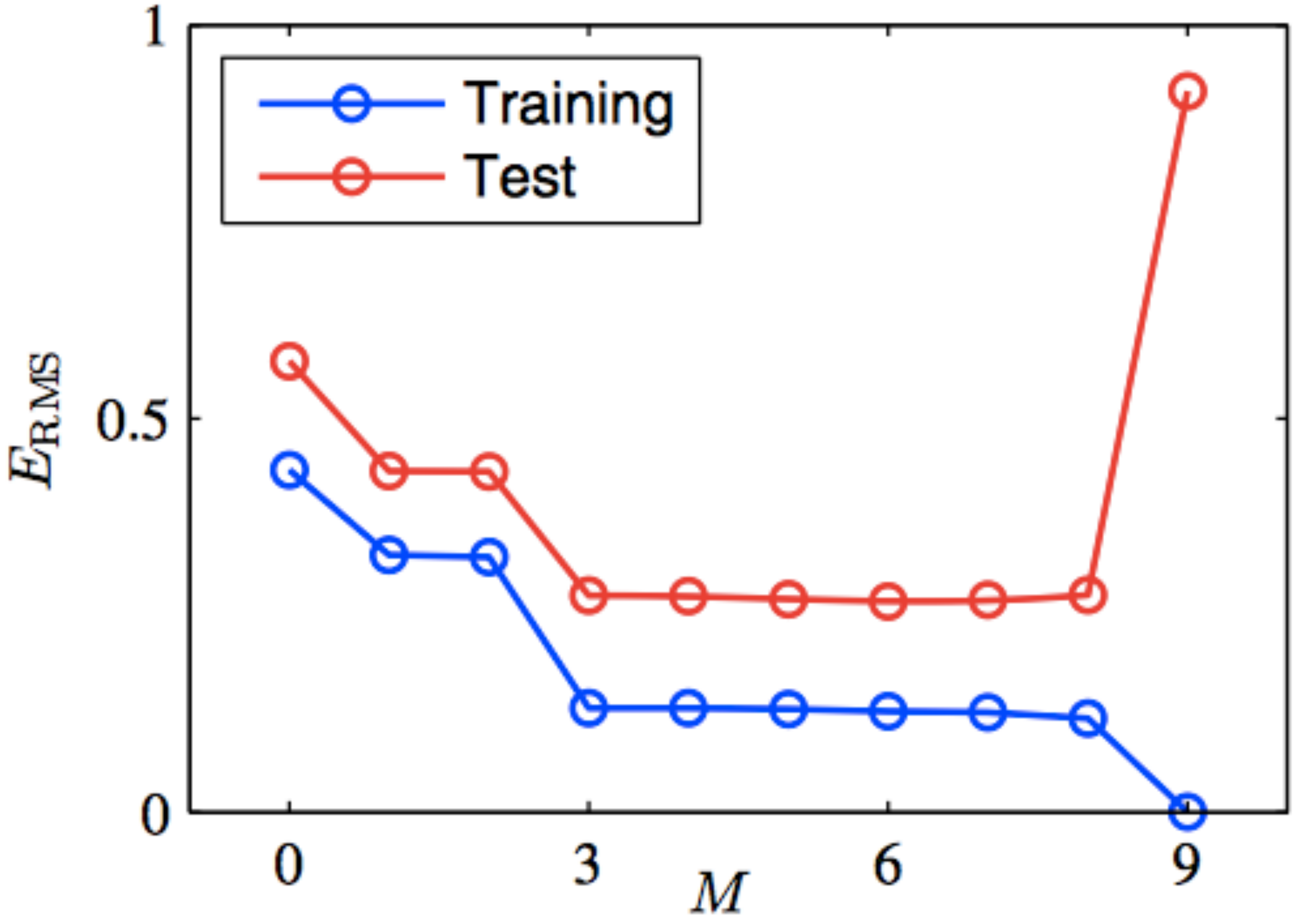


$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Coeficientes dos polinômios interpoladores

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

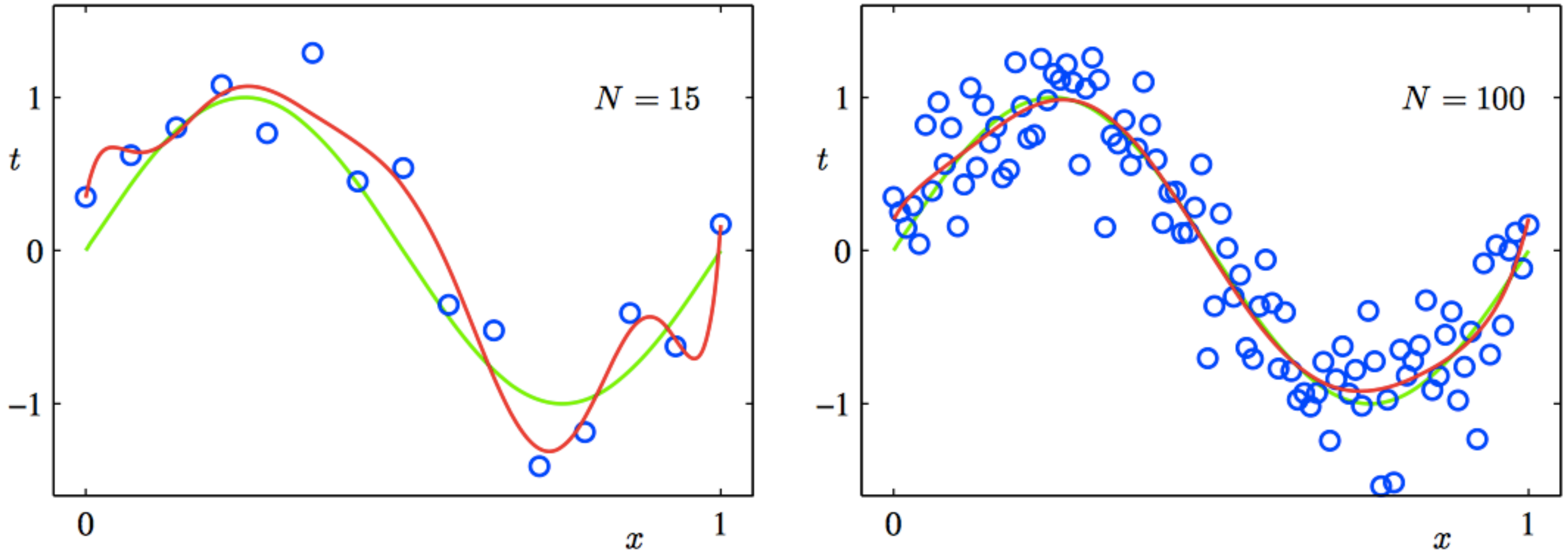
Overfitting (Treino é treino, jogo é jogo)



Conjunto de 100 pontos de teste.

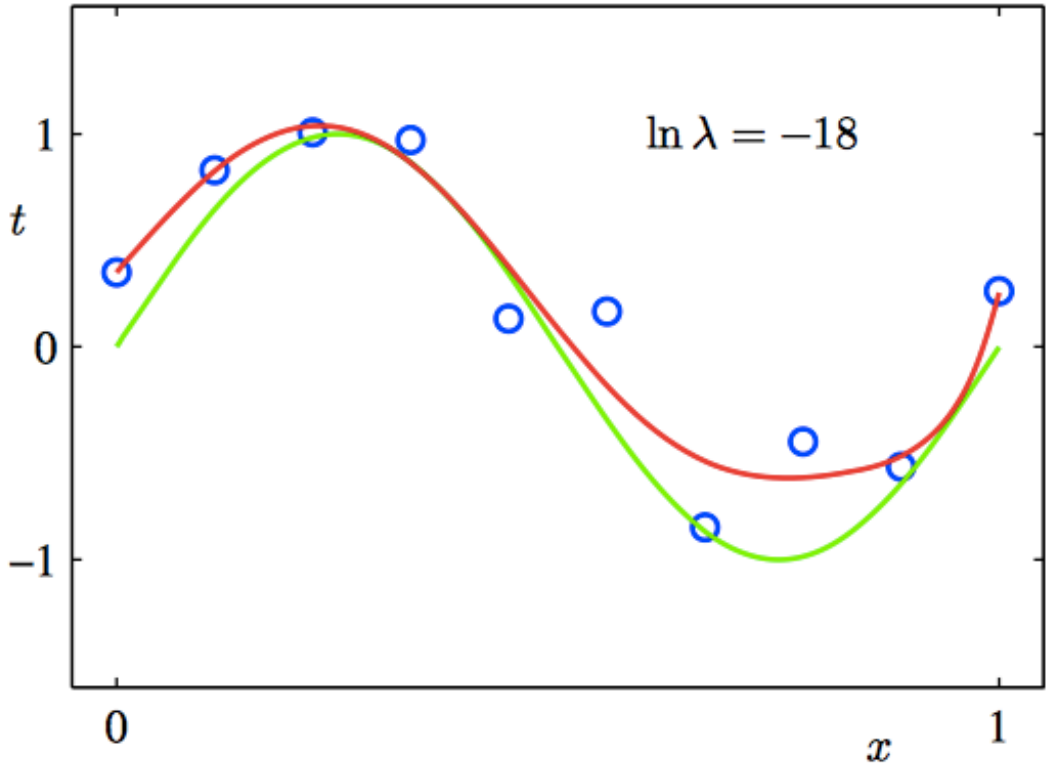
Treinando com mais dados

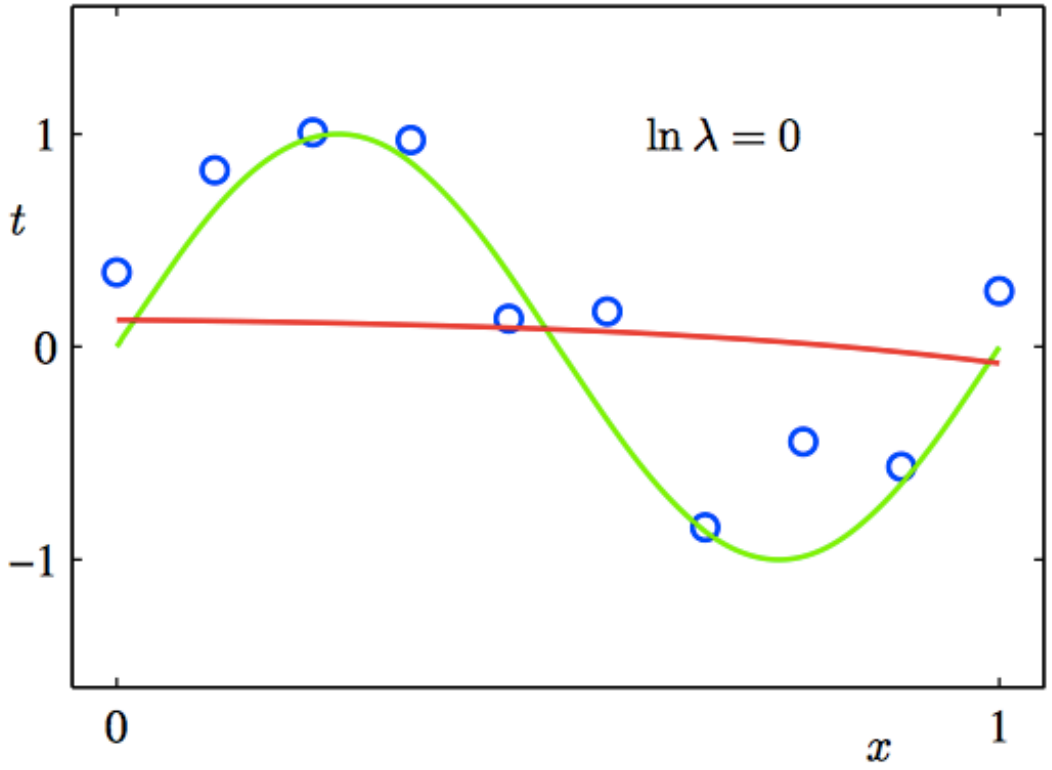
M=9



Penalização e Regularização

M=9





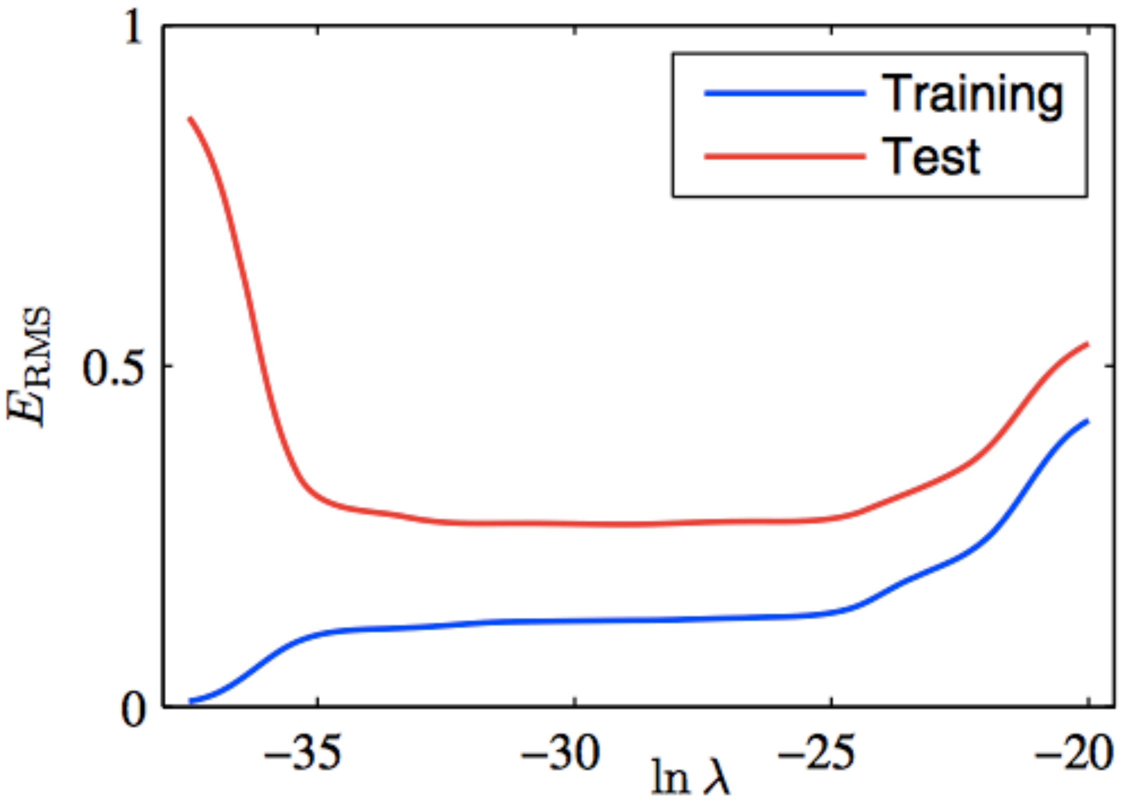
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Coeficientes dos polinômios interpoladores - o efeito regularizador

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

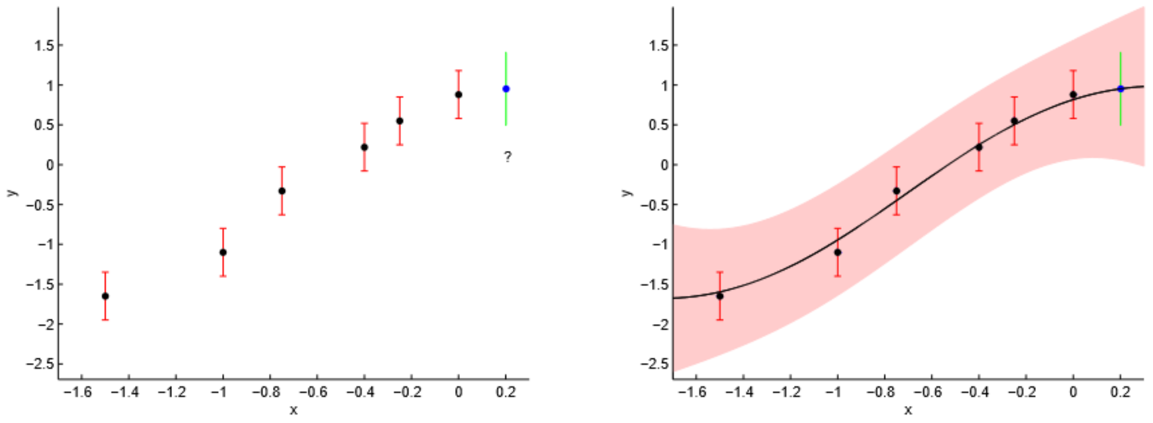
Domando o Overfitting (Equilibrando Treino e Jogo)

M=9



Uma abordagem bayesiana

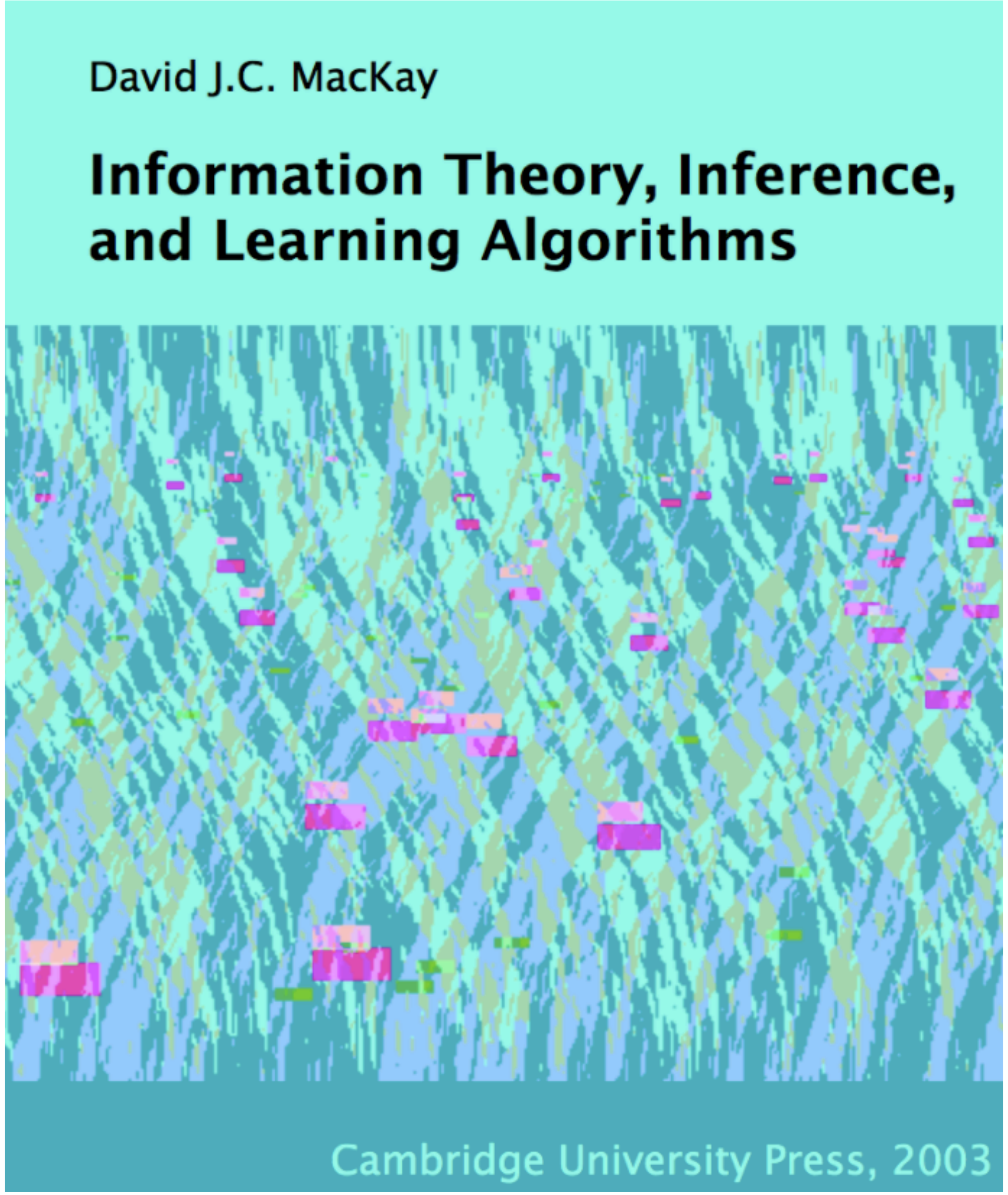
"Teoria da probabilidade não é nada além de senso comum reduzido ao cálculo" - Laplace



"Senso comum não é nada comum" - Voltaire

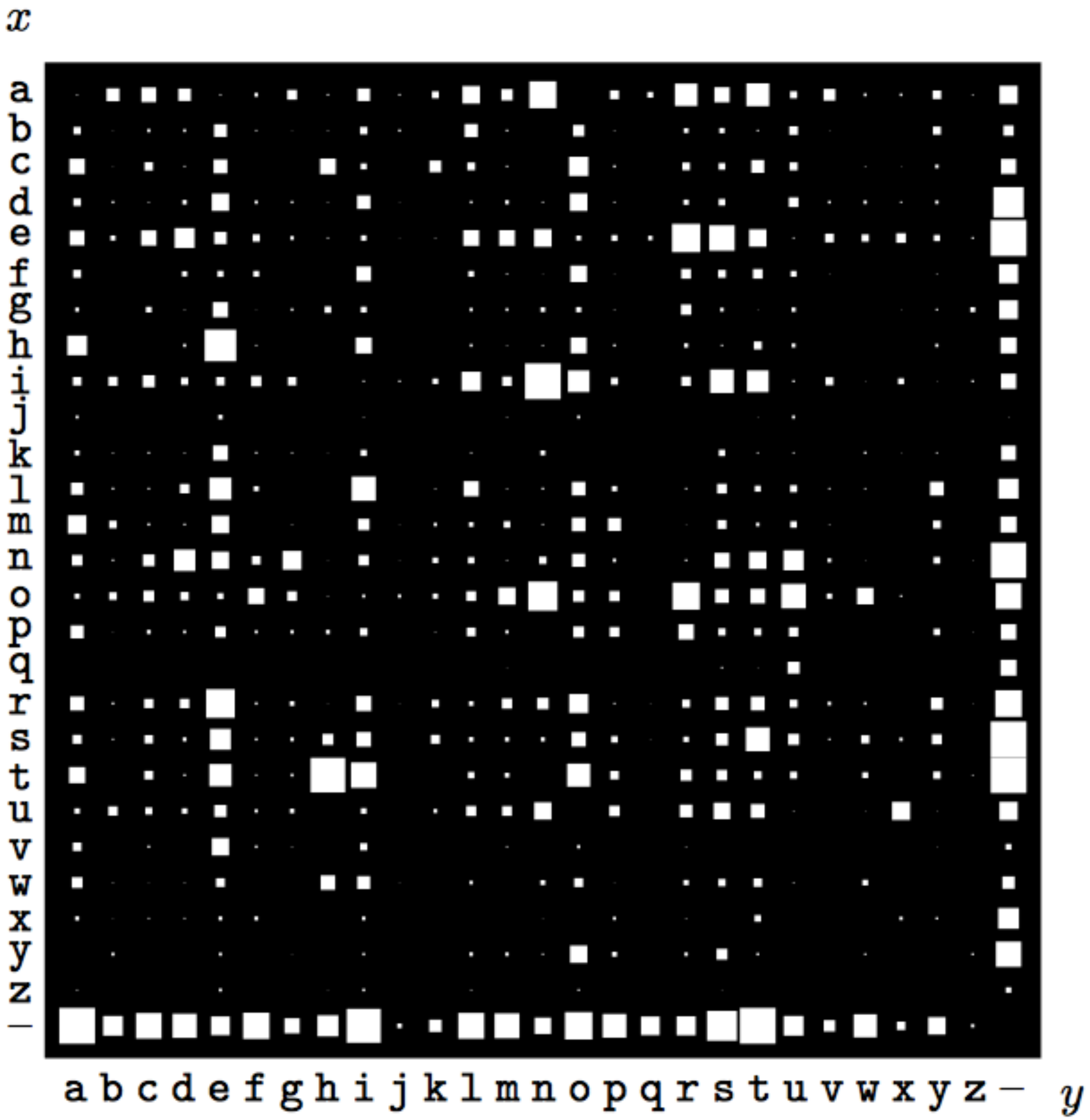
Sir David John Cameron MacKay

(22 de Abril 1967 – 14 de Abril 2016)



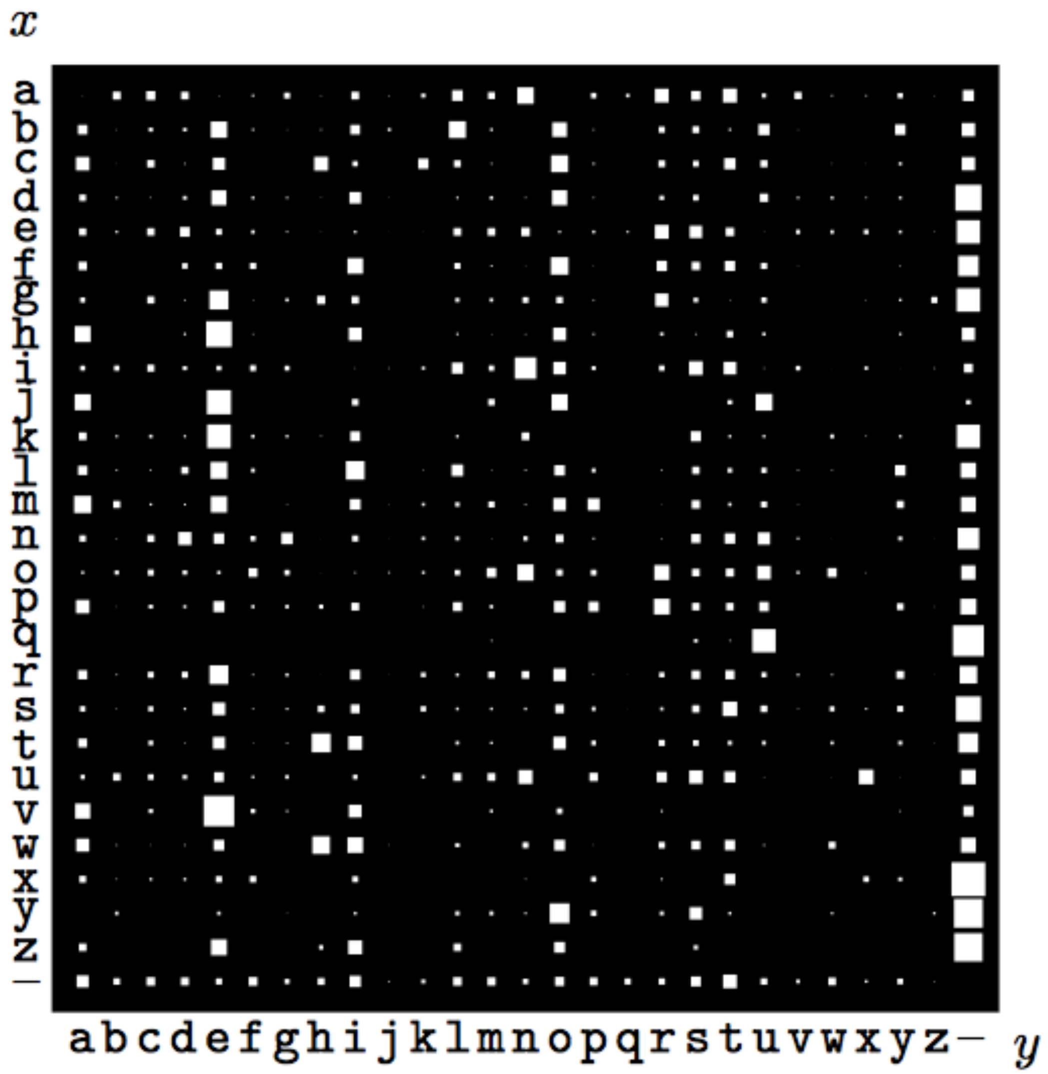
Teorema de Bayes

Busca por padrões – FAQ Linux

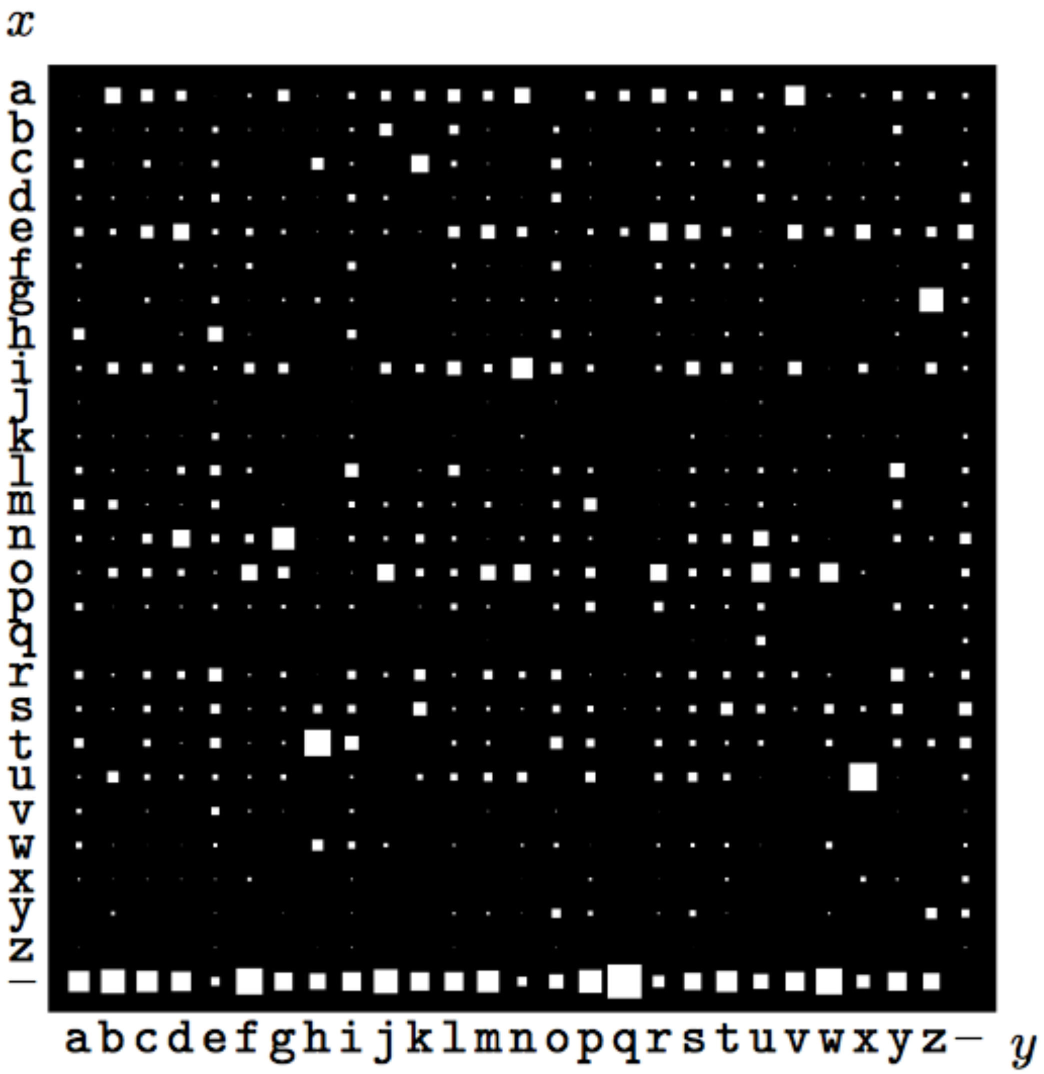


i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Teorema de Bayes



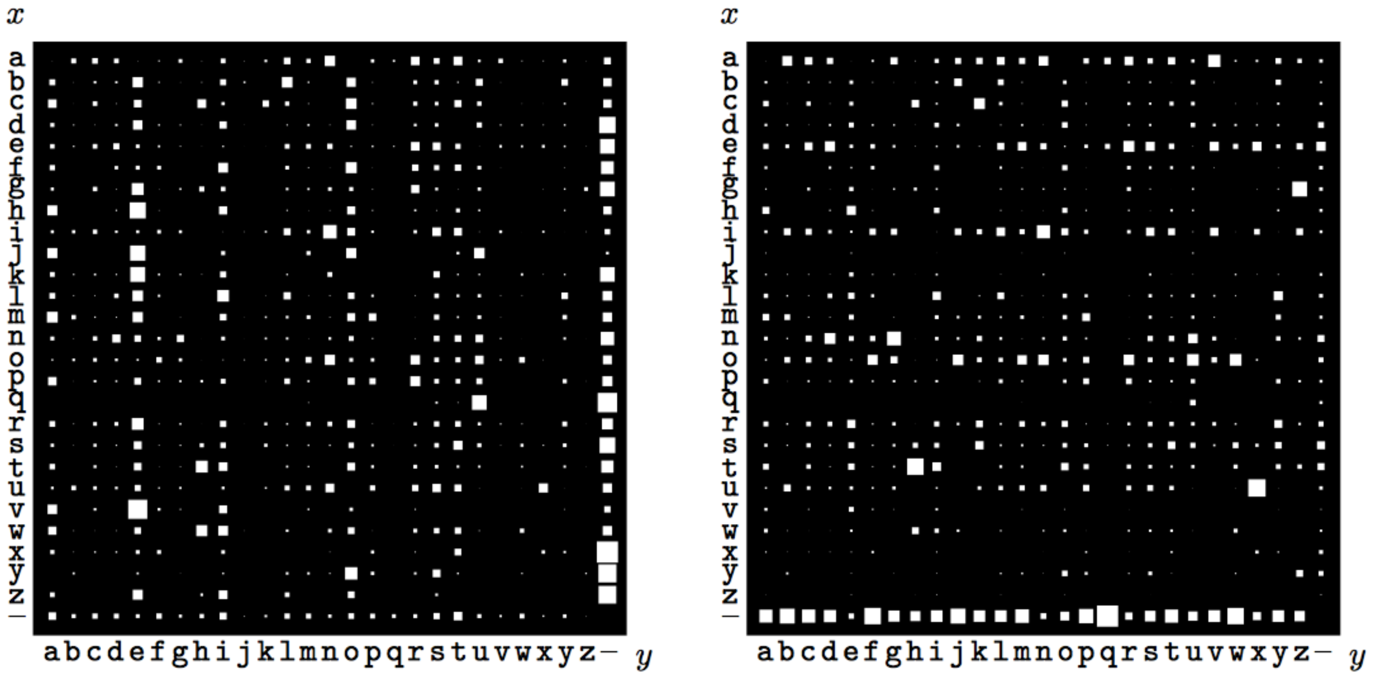
(a) $P(y \mid x)$



(b) $P(x \mid y)$

$$P(y \mid x, \mathcal{H}) = \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{P(x \mid \mathcal{H})}$$

$$= \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{\sum_{y'} P(x \mid y', \mathcal{H}) P(y' \mid \mathcal{H})}.$$



(a) $P(y \mid x)$ (b) $P(x \mid y)$

Previsão Bayesiana

0 preta 10 brancas	1 preta 9 brancas	....	9 pretas 1 branca	10 pretas 0 brancas

Urna u contém u bolas pretas e 10-u bolas brancas

Carlos seleciona uma urna ao acaso e retira N vezes com reposição

N=10 retiradas e $n_b = 3$ pretas foram observadas

Qual a probabilidade que Carlos esteja usando a urna u ?

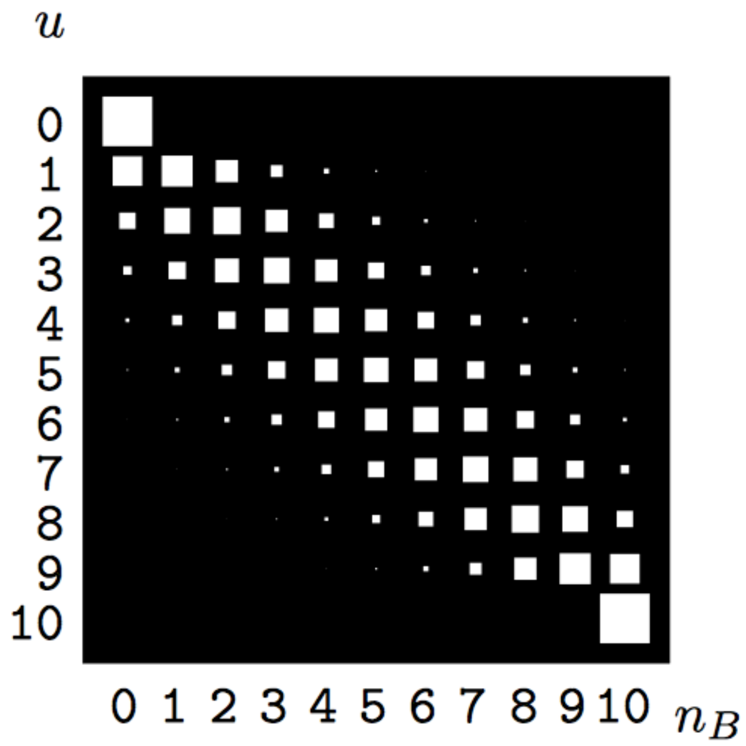
Inferência

0 preta 10 brancas	1 preta 9 brancas	....	9 pretas 1 branca	10 pretas 0 brancas

N=10 retiradas e $n_b = 3$ pretas foram observadas

$$
\begin{aligned}
P(u \mid n_B, N) &= \frac{P(u, n_B \mid N)}{P(n_B \mid N)} \\
&= \frac{P(n_B \mid u, N) P(u)}{P(n_B \mid N)}
\end{aligned}
$$

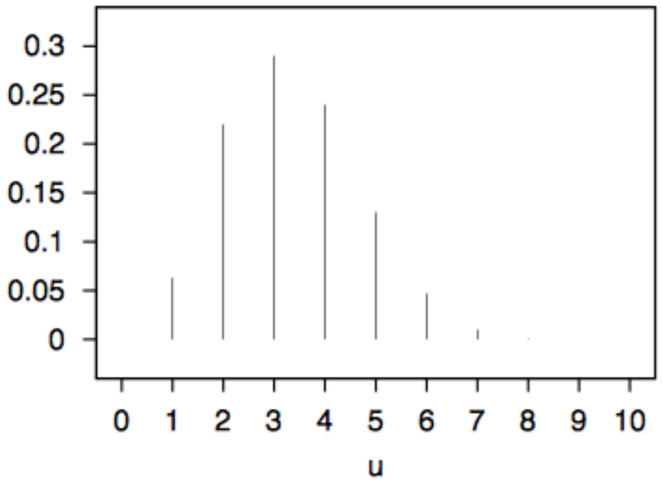
Você não pode fazer inferência sem fazer hipóteses



$$P(u) = \frac{1}{11}$$ Conhecimento a priori

$$P(n_B \,|\, u, N) = \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N-n_B}$$

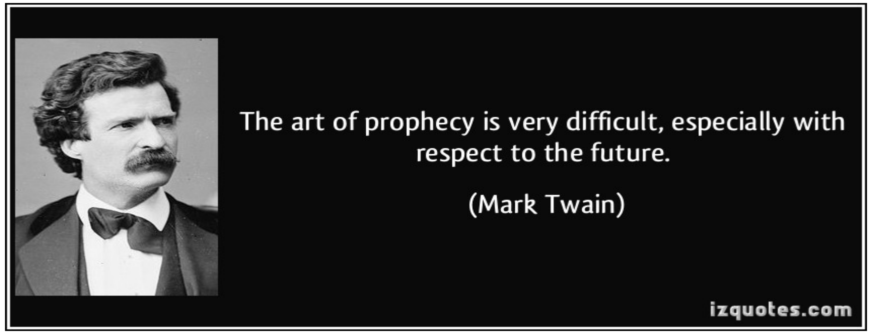
$$P(n_B \,|\, N) = \sum_u P(u, n_B \,|\, N) = \sum_u P(u) P(n_B \,|\, u, N)$$



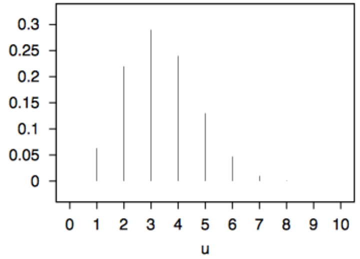
u	$P(u \mid n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.0099
8	0.00086
9	0.0000096
10	0

$$\begin{aligned} P(u \mid n_B, N) &= \frac{P(u)P(n_B \mid u, N)}{P(n_B \mid N)} \\ &= \frac{1}{P(n_B \mid N)} \frac{1}{11} \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N - n_B} \end{aligned}$$

Prediction



$$P(\text{ball}_{N+1} \text{ is black} \,|\, n_B, N) = \sum_u P(\text{ball}_{N+1} \text{ is black} \,|\, u, n_B, N) P(u \,|\, n_B, N)$$



u	$P(u \,\|\, n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.0099
8	0.00086
9	0.0000096
10	0

$$P(\text{ball}_{N+1} \text{ is black} \,|\, n_B = 3, N = 10) = 0.333$$

Previsão Bayesiana

$$p(y_*|\mathcal{D}) = \int p(y_*|\mathcal{D},\theta)\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}}\, d\theta$$

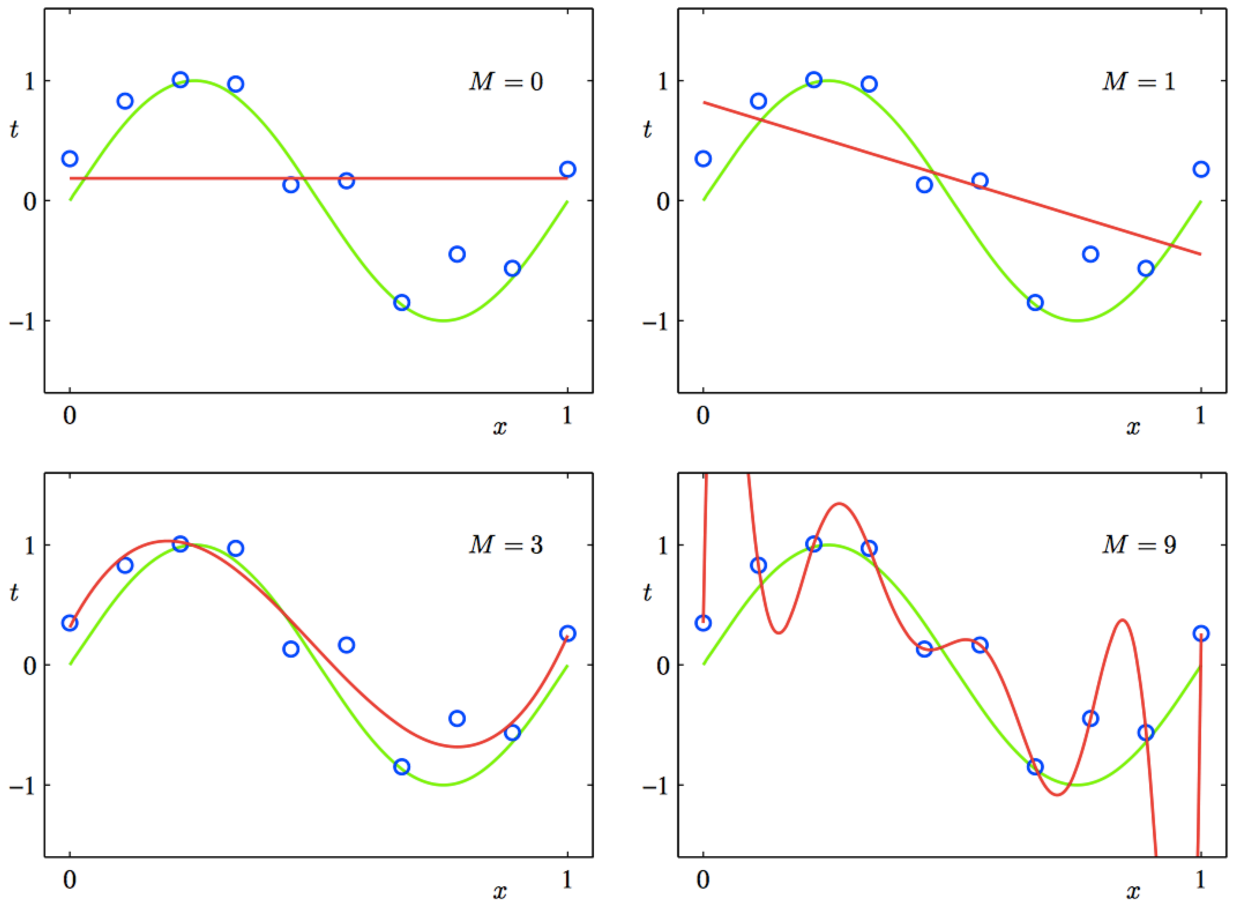
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta')p(\theta')\, d\theta'}$$

Bayes

$$p(y_*|\mathcal{D}) = \frac{\int p(y_*|\mathcal{D},\theta)p(\mathcal{D}|\theta)p(\theta)\, d\theta}{\int p(\mathcal{D}|\theta')p(\theta')\, d\theta'}$$

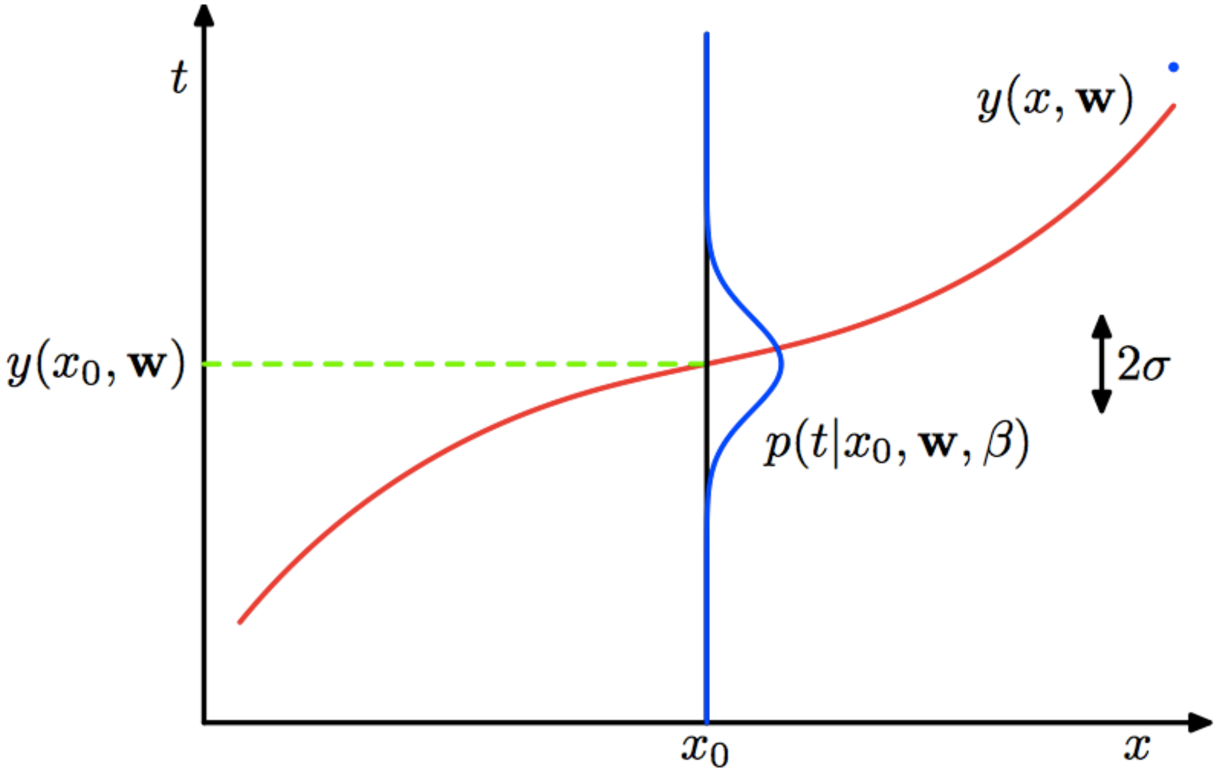
Custo exponencial com o número de parâmetros

Interpolação Polinomial – O Retorno



A abordagem bayesiana

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$ Modelo de Erro



$$N \text{ input values } \mathbf{x} = (x_1, \ldots, x_N)^T$$

$$\mathbf{t} \text{ target values } \mathbf{t} = (t_1, \ldots, t_N)^T$$

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

Independência dos Erros

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \boxed{\sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

Minimizando

$$\mathbf{w}_{\mathrm{ML}}$$

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t | y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right).$$

A abordagem bayesiana completa

Priori sobre os coeficientes

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathsf{x}, \mathsf{t}, \alpha, \beta) \propto p(\mathsf{t}|\mathsf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha).$$

MAP (Maximum a posteriori)

Equivale a Minimizar:

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \qquad \lambda = \frac{\alpha}{\beta}$$

Predição Bayesiana
(Tome médias levando em conta toda a incerteza)

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}.$$

$$\boxed{p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)}$$

Predição Bayesiana

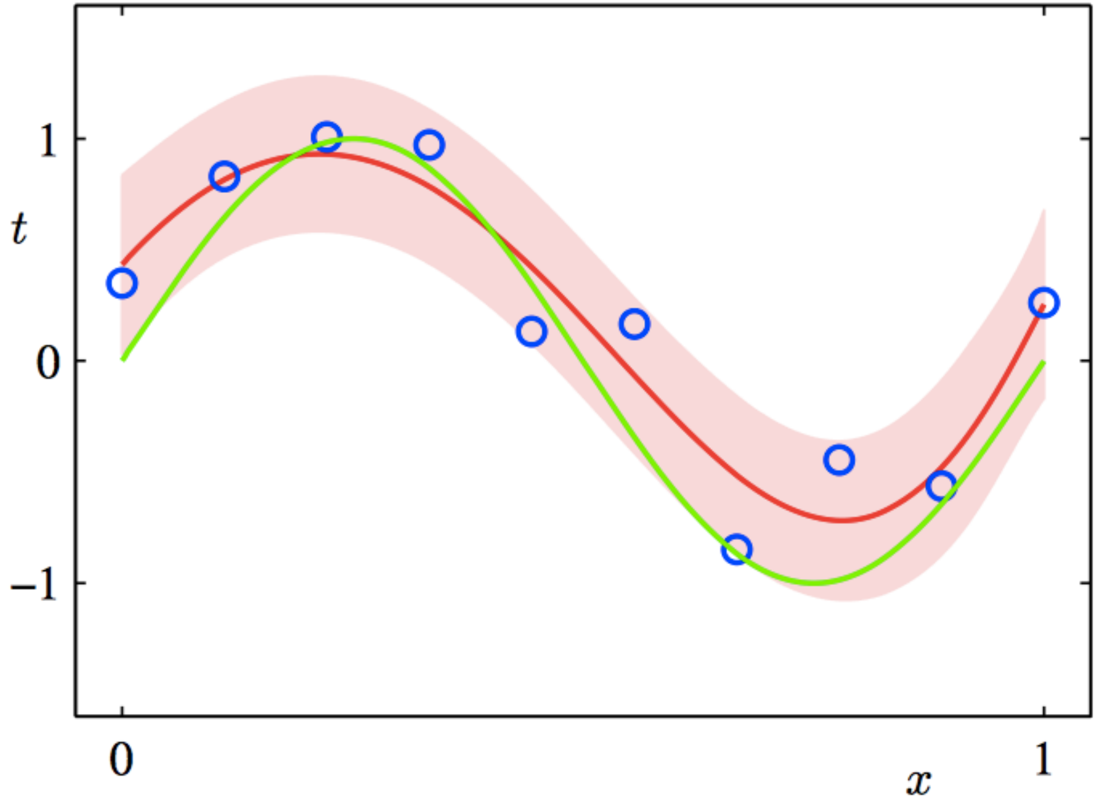
INCERTEZA DA OBSERVAÇÃO
+
INCERTEZA NOS PARÂMETROS
DO MODELO

$$\boxed{p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)}$$

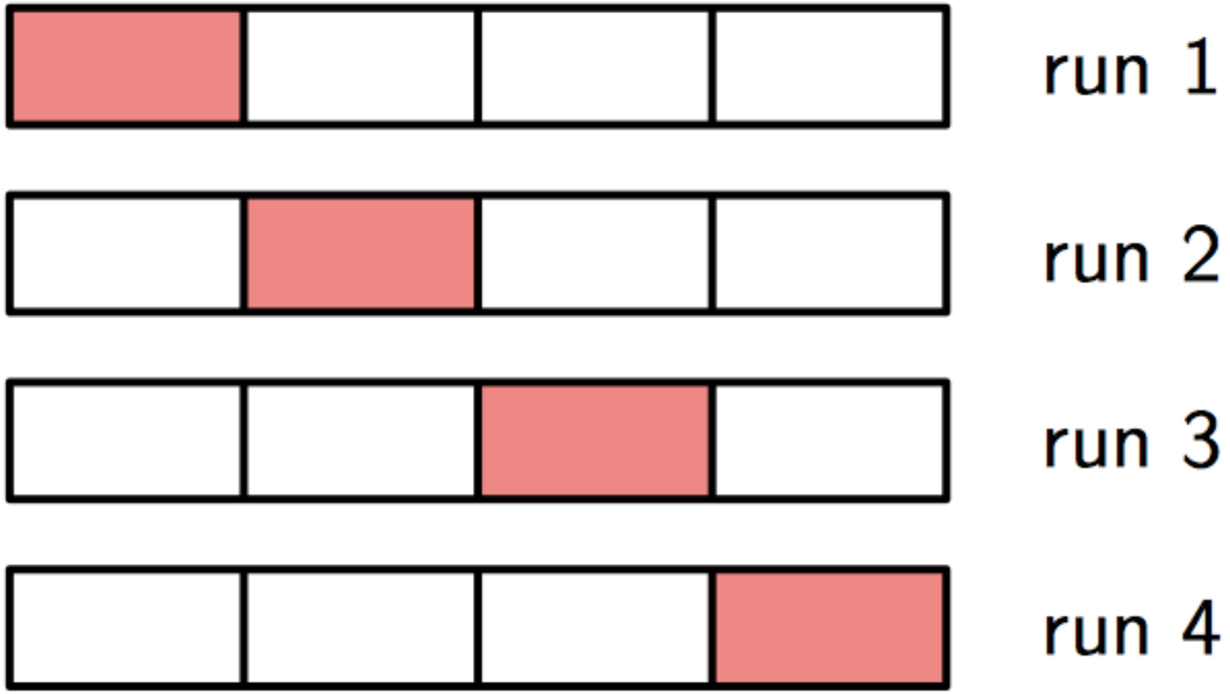
$$\alpha = 5 \times 10^{-3}$$

$$\beta = 11.1$$

Como escolher ?



Cross-validation and Model Choice



Posterior, likelihood, prior, evidence

$$P(\boldsymbol{\theta} \mid D, \mathcal{H}) = \frac{P(D \mid \boldsymbol{\theta}, \mathcal{H}) P(\boldsymbol{\theta} \mid \mathcal{H})}{P(D \mid \mathcal{H})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

A Navalha de Occam

Quais os dois próximos números ?

-1, 3, 7, 11, ?, ?

15, 19 ? (somar 4)

Ou

-19.9, 1043.8 ?

$$-x^3/11 + 9/11x^2 + 23/11\text{'?}$$

"A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data "– Paul Dirac

A Navalha de Occam

$\mathcal{H}_a$ – the sequence is an *arithmetic* progression, 'add n', where n is an integer.

~~$\mathcal{H}_c$ – the sequence is generated by a *cubic* function of the form $x \to cx^3 + dx^2 + e$, where c, d and e are fractions.~~

William of Occam

1287 – 1347

Entia non sunt multiplicanda sine necessitate

Lançamento de moedas – Seleção de Modelos

Observation of a sequence of the letters a and b

The probability, given p_a, that F tosses result in a sequence s that contains $\{F_a, F_b\}$ counts of the two outcomes

Probability of string **s**

$$\boxed{P(\mathbf{s} \mid p_{\mathrm{a}}, F, \mathcal{H}_1) = p_{\mathrm{a}}^{F_{\mathrm{a}}}(1-p_{\mathrm{a}})^{F_{\mathrm{b}}}.}$$

[For example, $P(\mathbf{s} = \mathtt{aaba} \mid p_{\mathrm{a}}, F=4, \mathcal{H}_1) = p_{\mathrm{a}} p_{\mathrm{a}} (1-p_{\mathrm{a}}) p_{\mathrm{a}}$.]

Priori

$$\boxed{P(p_{\mathrm{a}} \mid \mathcal{H}_1) = 1, \qquad p_{\mathrm{a}} \in [0,1]}$$

and $p_{\mathrm{b}} \equiv 1 - p_{\mathrm{a}}$.

Probabilidade posteriori de p_a

$$P(p_{\mathsf{a}} \mid \mathbf{s}, F, \mathcal{H}_1) \quad = \quad \frac{P(\mathbf{s} \mid p_{\mathsf{a}}, F, \mathcal{H}_1) P(p_{\mathsf{a}} \mid \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_1)}$$

<u>BAYES</u>

$$\boxed{P(p_{\mathsf{a}} \mid \mathbf{s}, F, \mathcal{H}_1) \quad = \quad \frac{p_{\mathsf{a}}^{F_{\mathsf{a}}}(1-p_{\mathsf{a}})^{F_{\mathsf{b}}}}{P(\mathbf{s} \mid F, \mathcal{H}_1)}}$$

<u>Constante Normalizadora = Evidência</u>

$$P(\mathbf{s} \mid F, \mathcal{H}_1) = \int_0^1 \mathrm{d}p_{\mathsf{a}}\, p_{\mathsf{a}}^{F_{\mathsf{a}}}(1-p_{\mathsf{a}})^{F_{\mathsf{b}}} = \frac{\Gamma(F_{\mathsf{a}}+1)\Gamma(F_{\mathsf{b}}+1)}{\Gamma(F_{\mathsf{a}}+F_{\mathsf{b}}+2)} = \frac{F_{\mathsf{a}}!F_{\mathsf{b}}!}{(F_{\mathsf{a}}+F_{\mathsf{b}}+1)!}$$

Do ajuste para a previsão

$$P(\mathbf{a} \,|\, \mathbf{s}, F) \;=\; \int \mathrm{d}p_{\mathrm{a}}\, P(\mathbf{a} \,|\, p_{\mathrm{a}}) P(p_{\mathrm{a}} \,|\, \mathbf{s}, F)$$

$$P(\mathbf{a} \,|\, \mathbf{s}, F) = \int \mathrm{d}p_{\mathrm{a}}\, p_{\mathrm{a}} \frac{p_{\mathrm{a}}^{F_{\mathrm{a}}}(1-p_{\mathrm{a}})^{F_{\mathrm{b}}}}{P(\mathbf{s} \,|\, F)}$$

$$= \int \mathrm{d}p_{\mathrm{a}}\, \frac{p_{\mathrm{a}}^{F_{\mathrm{a}}+1}(1-p_{\mathrm{a}})^{F_{\mathrm{b}}}}{P(\mathbf{s} \,|\, F)}$$

$$= \left[\frac{(F_{\mathrm{a}}+1)!\, F_{\mathrm{b}}!}{(F_{\mathrm{a}}+F_{\mathrm{b}}+2)!}\right] \Big/ \left[\frac{F_{\mathrm{a}}!\, F_{\mathrm{b}}!}{(F_{\mathrm{a}}+F_{\mathrm{b}}+1)!}\right] \;=\; \frac{F_{\mathrm{a}}+1}{F_{\mathrm{a}}+F_{\mathrm{b}}+2}$$

REGRA DE LAPLACE

Seleção de modelos

Outro cientista introduz um outro modelo:

$$\text{Hipothesis } H_0\text{: probability of } a \text{ is } p_0 = \frac{1}{6}$$

$$P(\mathcal{H}_1 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1)}{P(\mathbf{s} \,|\, F)}.$$

Como selecionar o modelo ?

vs.

$$P(\mathcal{H}_0 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)}{P(\mathbf{s} \,|\, F)}$$

$$P(\mathbf{s} \,|\, F) = P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1) + P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)$$

Seleção de modelos

Defining p_0 to be 1/6, we have

$$P(\mathbf{s} \mid F, \mathcal{H}_0) = p_0^{F_\mathtt{a}}(1 - p_0)^{F_\mathtt{b}}$$ = Evidência

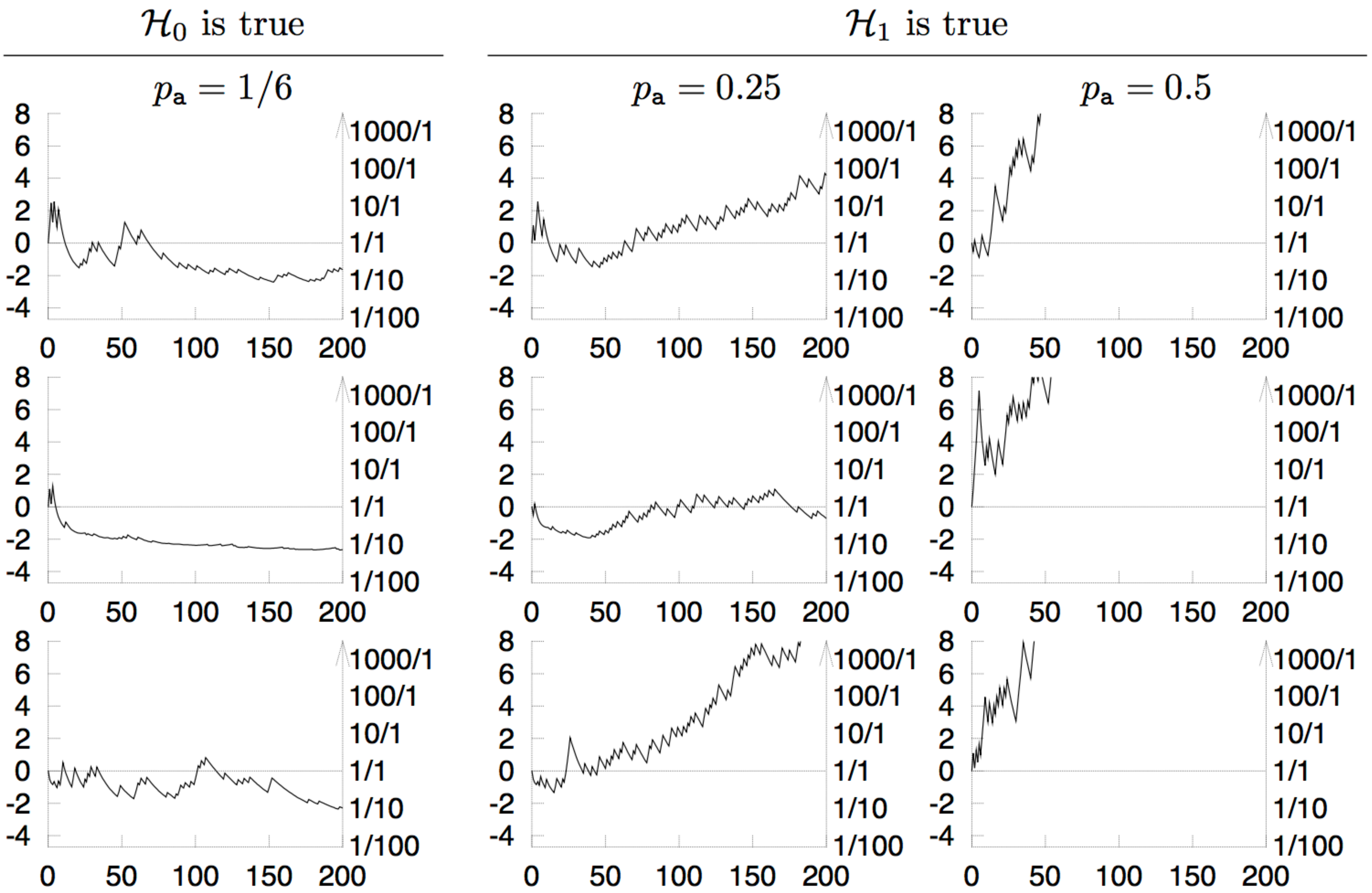
Thus the posterior probability ratio of model $\mathcal{H}_1$ to model $\mathcal{H}_0$ is

$$\begin{aligned}\frac{P(\mathcal{H}_1 \mid \mathbf{s}, F)}{P(\mathcal{H}_0 \mid \mathbf{s}, F)} &= \frac{P(\mathbf{s} \mid F, \mathcal{H}_1)P(\mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)P(\mathcal{H}_0)} \\ &= \frac{F_\mathtt{a}!F_\mathtt{b}!}{(F_\mathtt{a} + F_\mathtt{b} + 1)!} \Big/ p_0^{F_\mathtt{a}}(1 - p_0)^{F_\mathtt{b}}.\end{aligned}$$

How model comparison works: The evidence for a model is usually the normalizing constant of an earlier Bayesian inference.

Comparando modelos

F	Data $(F_{\mathtt{a}}, F_{\mathtt{b}})$	$\dfrac{P(\mathcal{H}_1 \mid \mathbf{s}, F)}{P(\mathcal{H}_0 \mid \mathbf{s}, F)}$	
6	$(5, 1)$	222.2	
6	$(3, 3)$	2.67	
6	$(2, 4)$	0.71	$= 1/1.4$
6	$(1, 5)$	0.356	$= 1/2.8$
6	$(0, 6)$	0.427	$= 1/2.3$
20	$(10, 10)$	96.5	
20	$(3, 17)$	0.2	$= 1/5$
20	$(0, 20)$	1.83	



$\mathcal{H}_0$ is true
$\mathcal{H}_1$ is true
$p_a = 1/6$
$p_a = 0.25$
$p_a = 0.5$
8 6 4 2 0 -2 -4
1000/1 100/1 10/1 1/1 1/10 1/100
0 50 100 150 200

Navalha de Occam – Ajuste de modelo

1 – First level of inference

$$P(\mathbf{w}|D,\mathcal{H}_i) = \frac{P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)}{P(D|\mathcal{H}_i)}$$

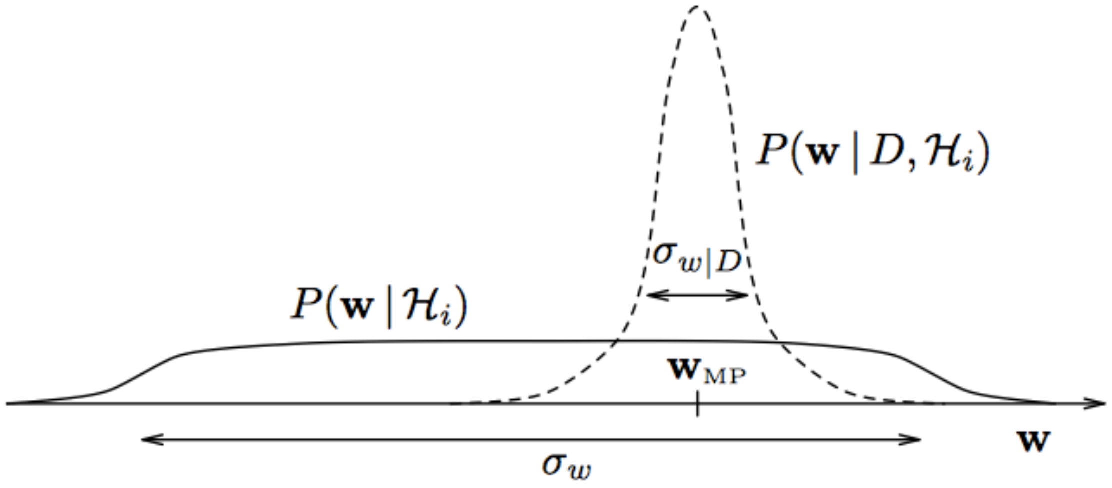
$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$$

Navalha de Occam – Comparação de modelos

2 – Second level of inference

$$P(\mathcal{H}_i \,|\, D) \propto P(D \,|\, \mathcal{H}_i) P(\mathcal{H}_i)$$

$$\boxed{P(D \,|\, \mathcal{H}_i) = \int P(D \,|\, \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} \,|\, \mathcal{H}_i) \, \mathrm{d}\mathbf{w}}$$

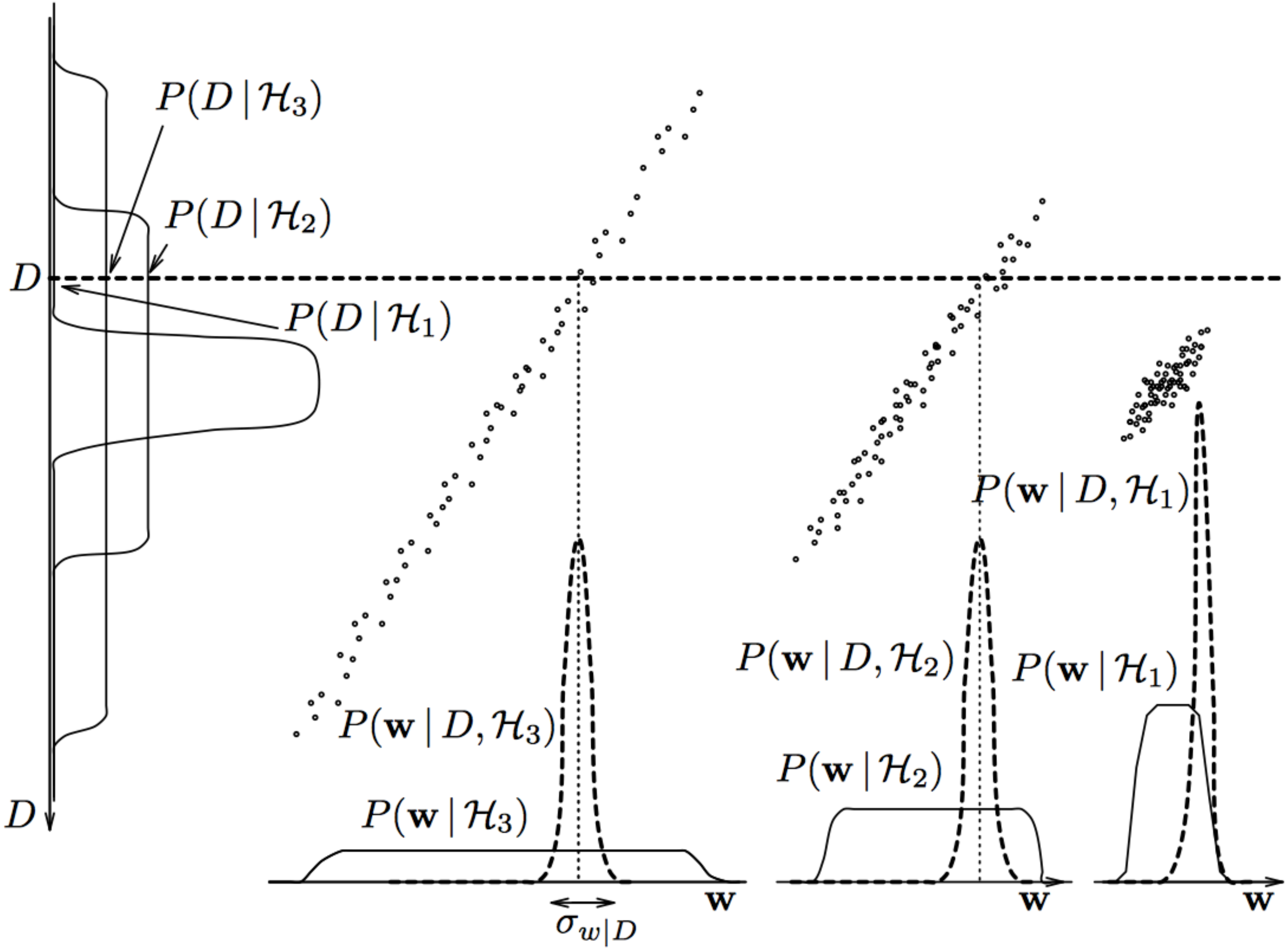


For many problems the posterior $P(\mathbf{w} \,|\, D, \mathcal{H}_i) \propto P(D \,|\, \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} \,|\, \mathcal{H}_i)$ has a strong peak at the most probable parameters $\mathbf{w}_{\mathrm{MP}}$

$$P(\mathbf{w}_{\mathrm{MP}} \,|\, \mathcal{H}_i) = 1/\sigma_w$$

$$\underset{\text{Evidence}}{P(D \,|\, \mathcal{H}_i)} \simeq \underbrace{P(D \,|\, \mathbf{w}_{\mathrm{MP}}, \mathcal{H}_i)}_{\text{Best fit likelihood}} \times \underbrace{P(\mathbf{w}_{\mathrm{MP}} \,|\, \mathcal{H}_i) \, \sigma_{w|D}}_{\text{Occam factor}}$$

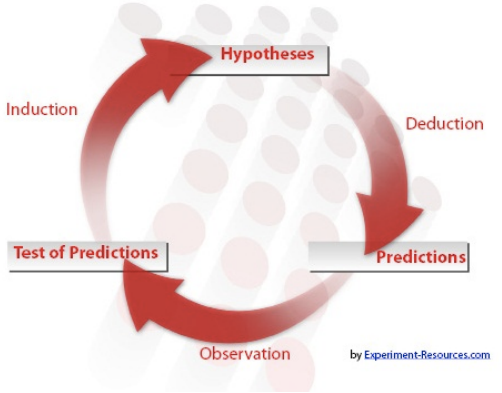
$$\boxed{\text{Occam factor} = \frac{\sigma_{w|D}}{\sigma_w}}$$


$P(D \mid \mathcal{H}_3)$
$P(D \mid \mathcal{H}_2)$
D
$P(D \mid \mathcal{H}_1)$
$P(\mathbf{w} \mid D, \mathcal{H}_1)$
$P(\mathbf{w} \mid D, \mathcal{H}_2)$
$P(\mathbf{w} \mid \mathcal{H}_1)$
$P(\mathbf{w} \mid D, \mathcal{H}_3)$
$P(\mathbf{w} \mid \mathcal{H}_2)$
D
$P(\mathbf{w} \mid \mathcal{H}_3)$
$\mathbf{w}$
$\mathbf{w}$
$\mathbf{w}$
$\sigma_{w|D}$

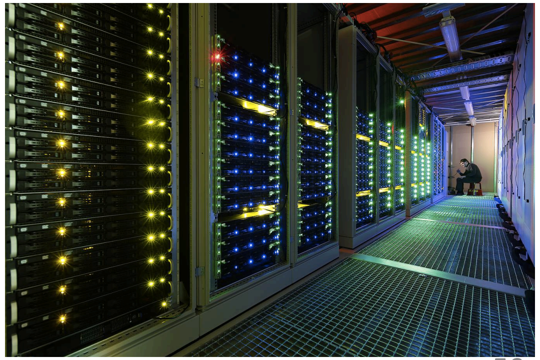
Probability theory reaches parts that ad hoc methods cannot reach

$$P(\mathbf{t} \mid D, I) = \sum_{\mathcal{H}} P(\mathbf{t} \mid D, \mathcal{H}, I) P(\mathcal{H} \mid D, I).$$

Previsão Bayesiana



$$P(\mathcal{H} \mid D, I) = \frac{P(D \mid \mathcal{H}, I) P(\mathcal{H} \mid I)}{P(D \mid I)}$$



$$P(\lambda \mid D, \mathcal{H}) = \frac{P(D \mid \lambda, \mathcal{H}) P(\lambda \mid \mathcal{H})}{\underline{P(D \mid \mathcal{H})}}.$$

EVIDENCE

Machine Learning: Treino, Teste, Previsão e Ação