Learning from Observations

Bishop, Ch.1

Russell & Norvig Ch. 18

Learning as source of knowledge

- **Implicit models**: In many domains, we cannot say how we manage to perform so well
- **Unknown environment:** After some effort, we can get a system to work for a finite environment, but it fails in new areas
- **Model structures**: Learning can reveal properties (regularities) of the system behaviour
  - Modifies agent's decision models to **reduce complexity** and improve performance

Feedback in Learning

- Type of feedback:
  - Supervised learning: correct answers for each example
    - Discrete (categories) : classification
    - Continuous : regression
  - Unsupervised learning: correct answers not given
  - Reinforcement learning: occasional rewards

Inductive learning

- Simplest form: learn a function from examples

An example is a pair (x, y) : x = data, y = outcome
assume: y drawn from function f(x) : y = f(x) + noise

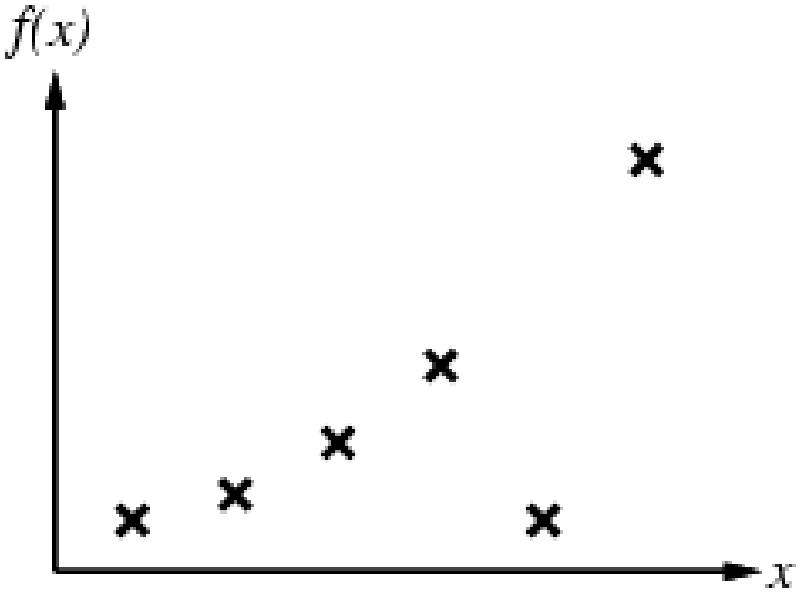
f = target function

Problem: find a hypothesis h
such that $h \approx f$
given a training set of examples

Note: highly simplified model :
- Ignores prior knowledge : some h may be more likely
- Assumes lots of examples are available
- Objective: maximize prediction for unseen data – Q. How?

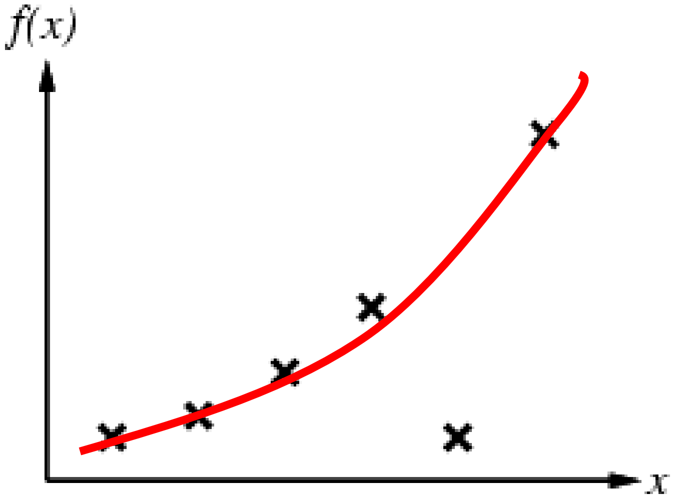
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



Regression vs Classification

$y = f(x)$



Regression:

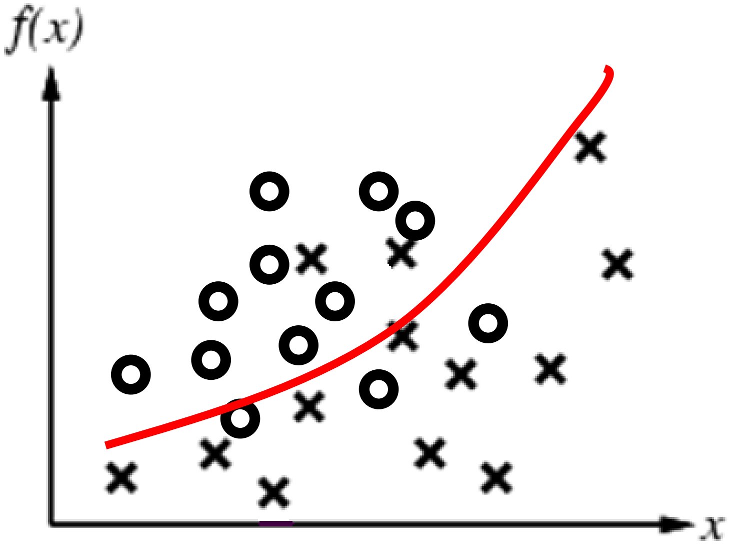
y is continuous

Classification:

y : set of discrete values

e.g. classes C_1, C_2, C_3...

$y \in \{1,2,3...\}$



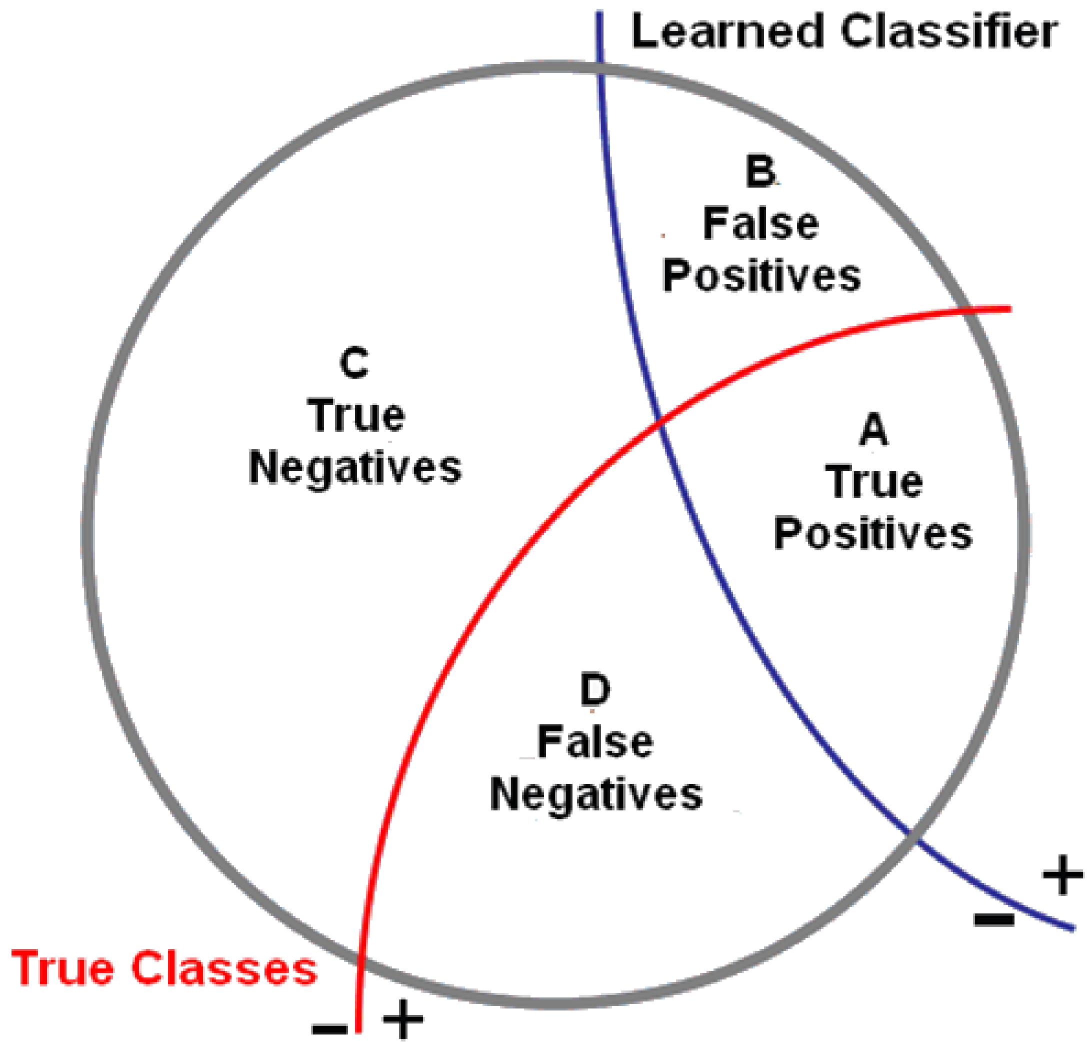
Precision vs Recall

Precision:

A / Retrieved Positives

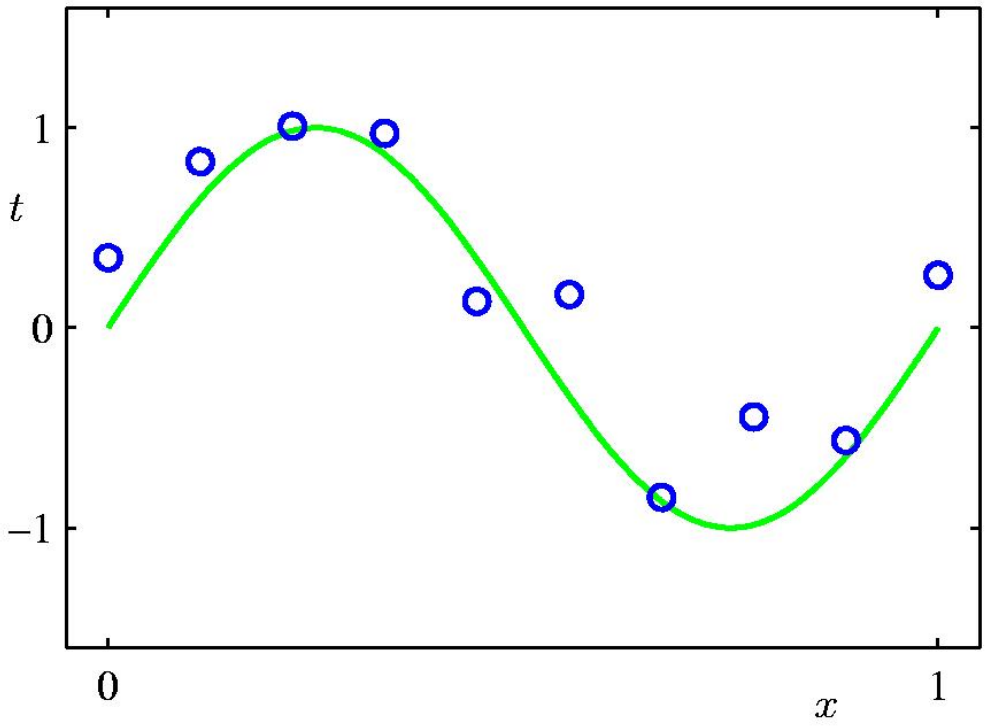
Recall:

A / Actual Positives



Regression

Polynomial Curve Fitting



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

Linear Regression

$$\mathbf{y = f(x) = \Sigma_i\, w_i \,.\, \varphi_i(x)}$$

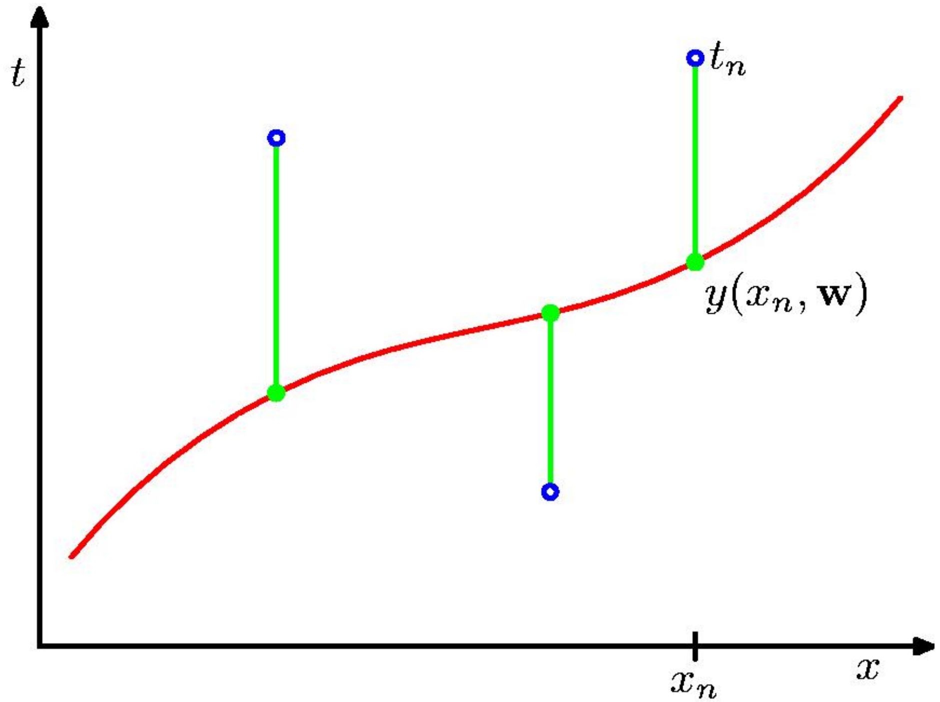
$\boldsymbol{\varphi_i(x)}$: basis function

w_i : weights

Linear : function is linear in the weights

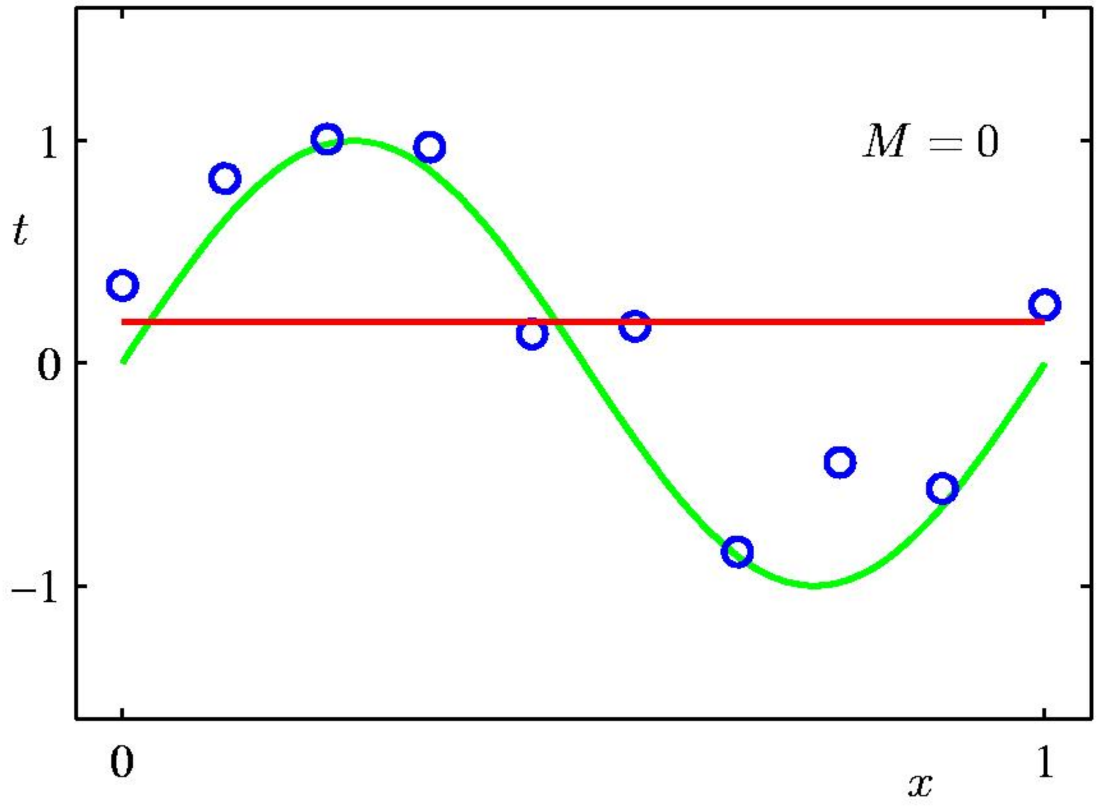
Quadratic error function --> derivative is linear in **w**

Sum-of-Squares Error Function

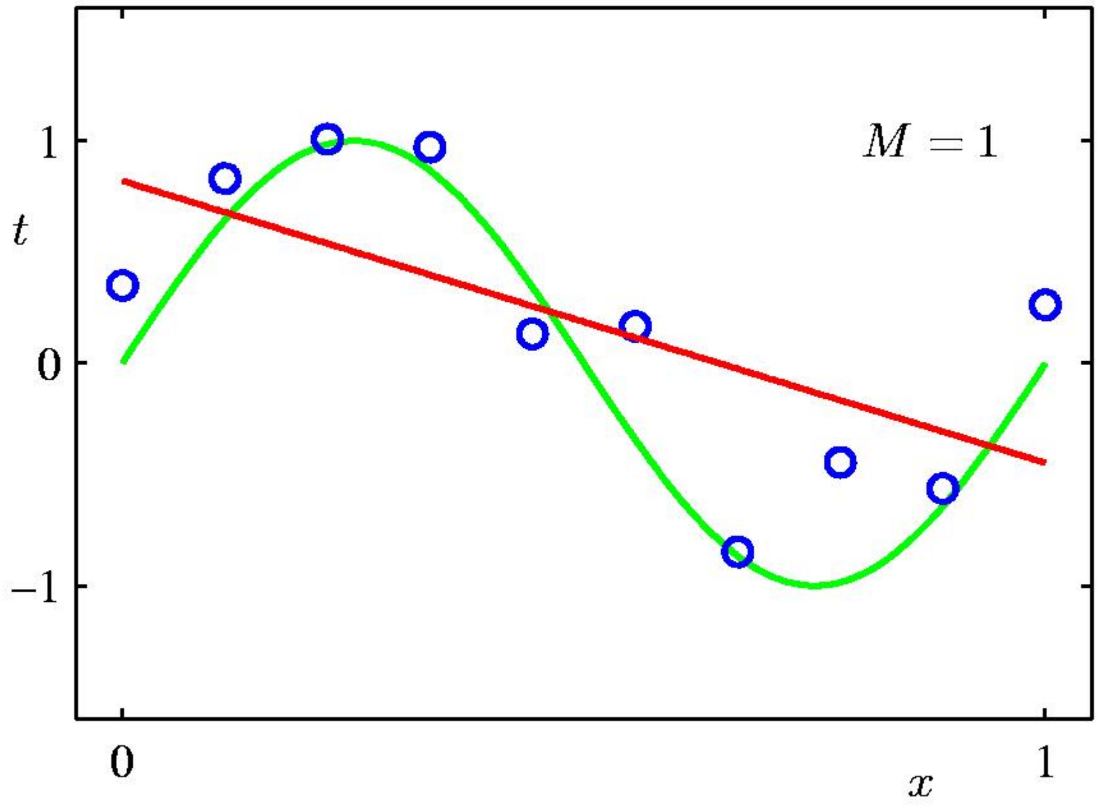


$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

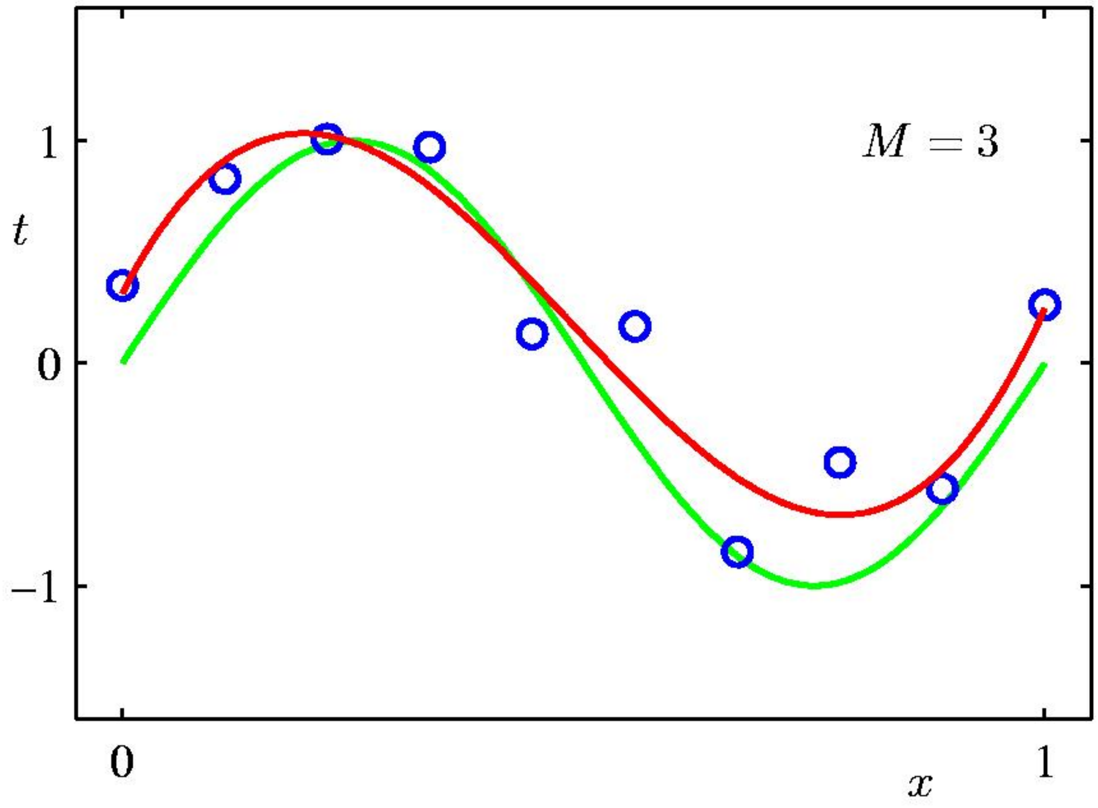
0^{th} Order Polynomial



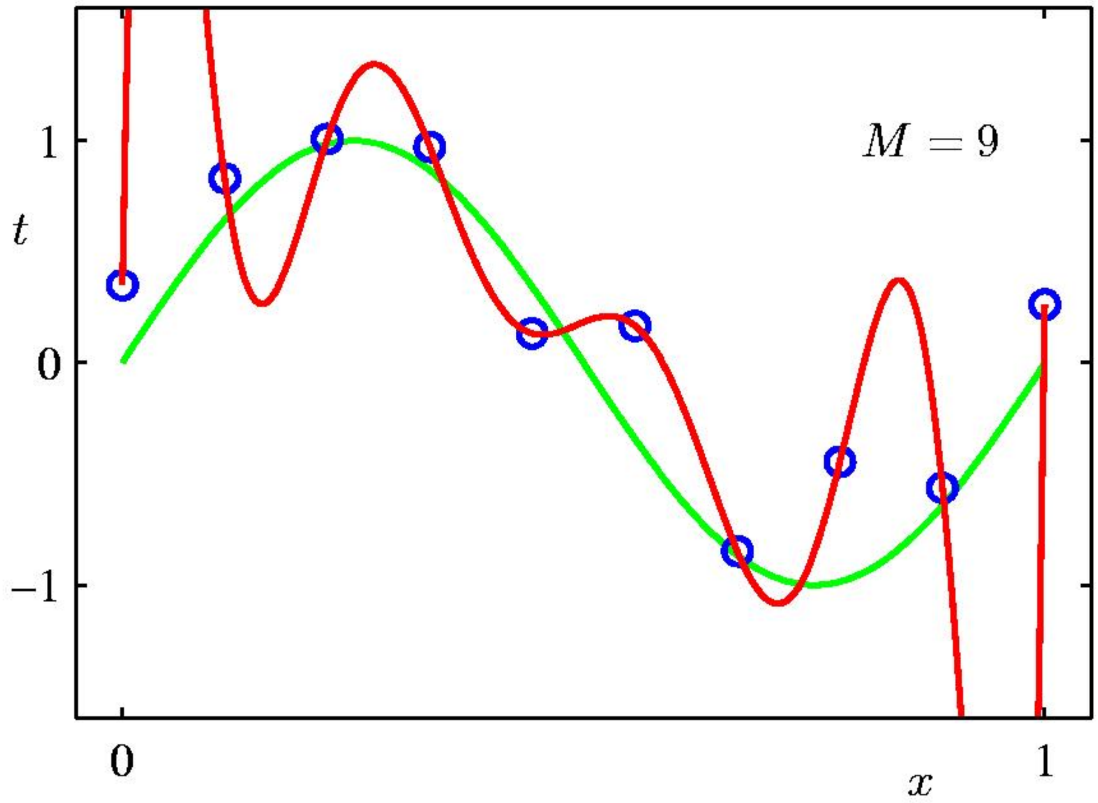
1^{st} Order Polynomial



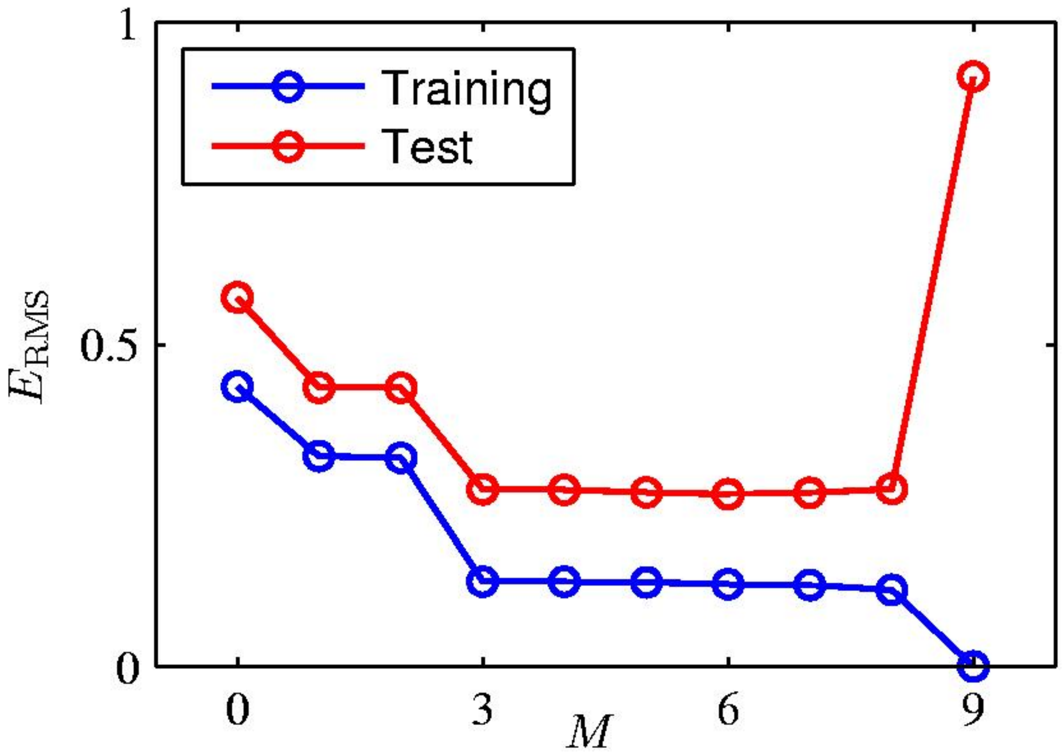
3rd Order Polynomial



9^{th} Order Polynomial



Over-fitting

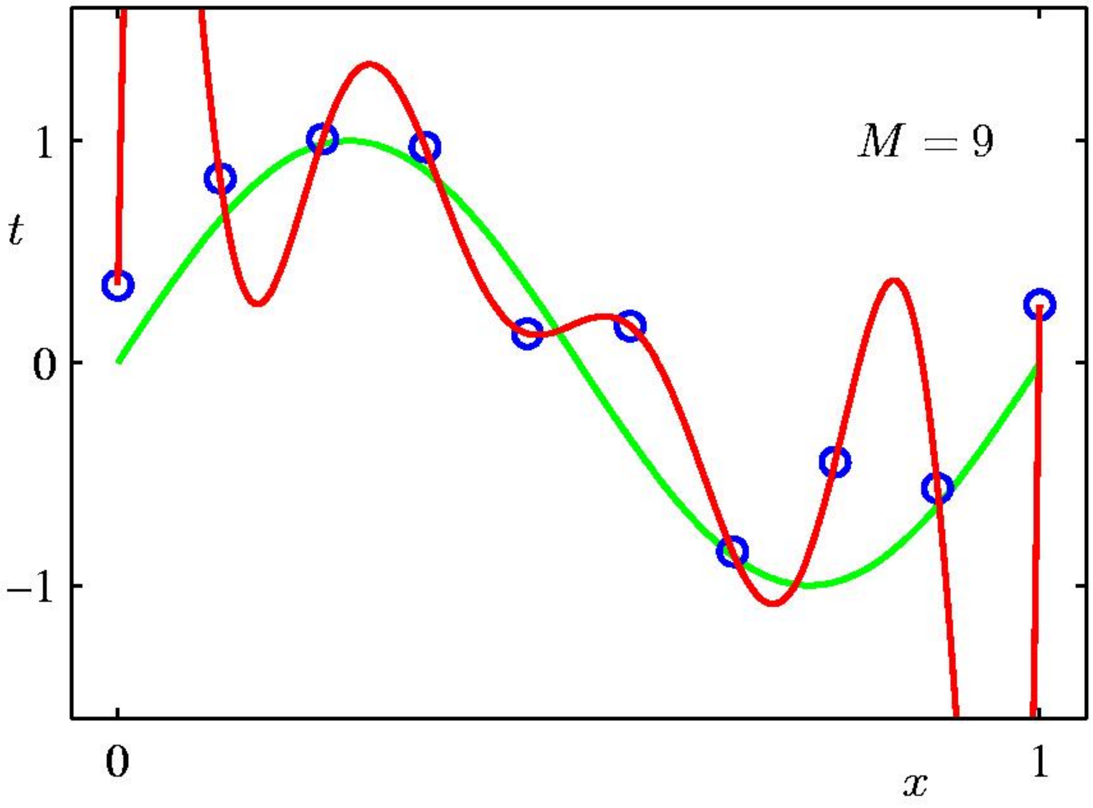


Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

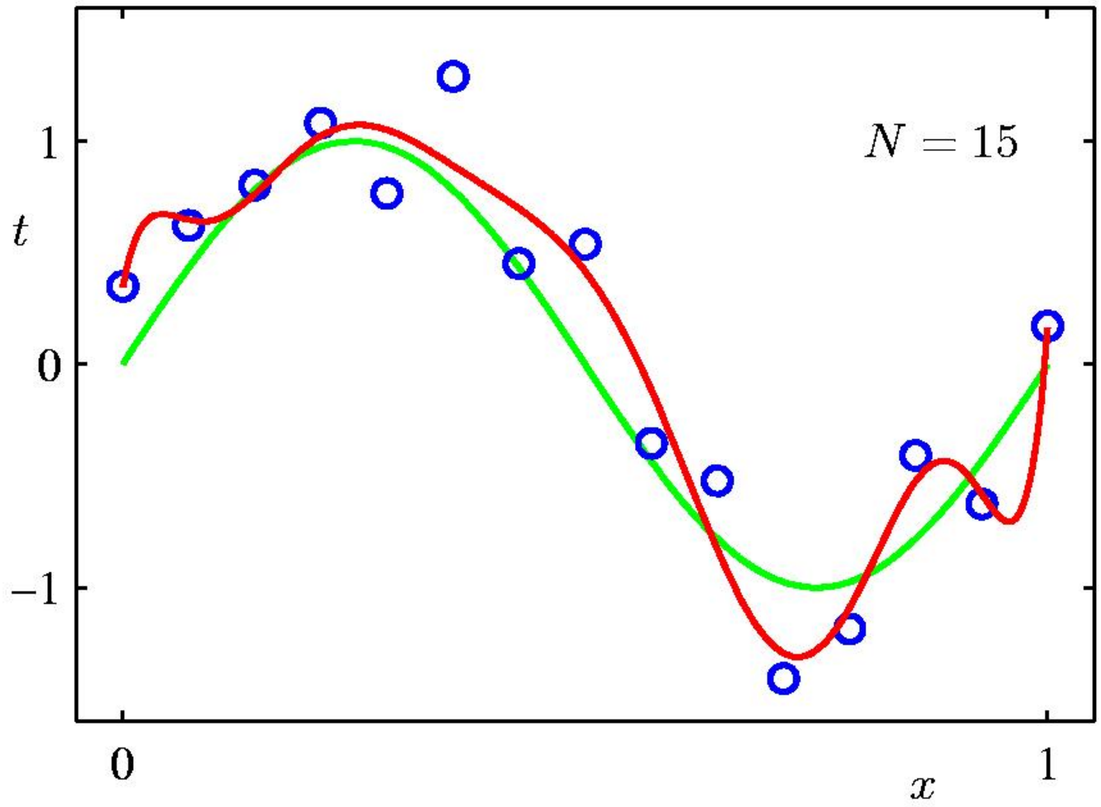
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

9^{th} Order Polynomial



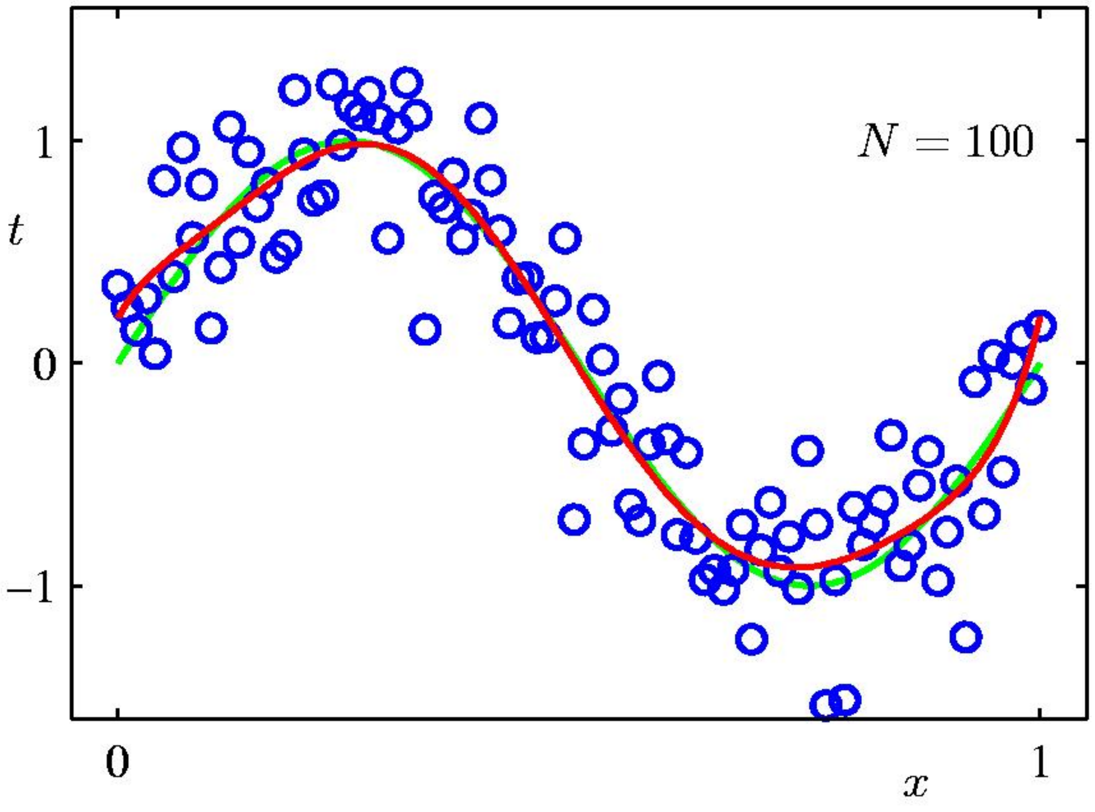
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

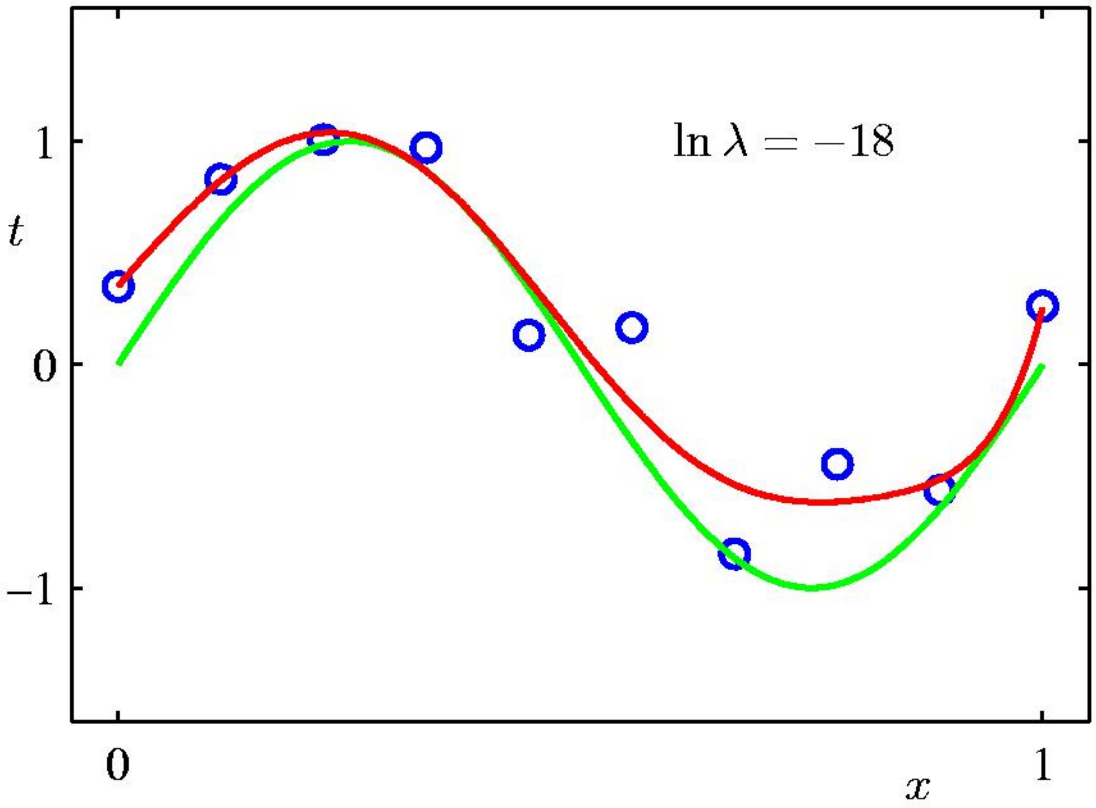


Regularization

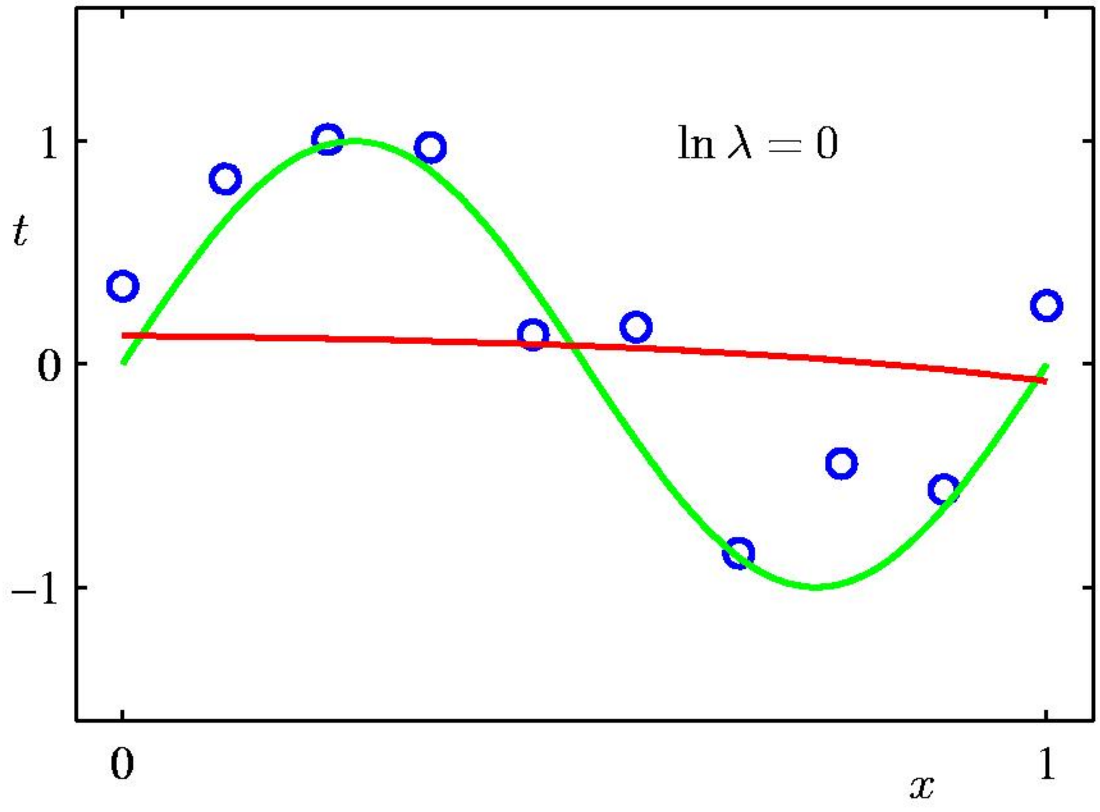
Penalize large coefficient values

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

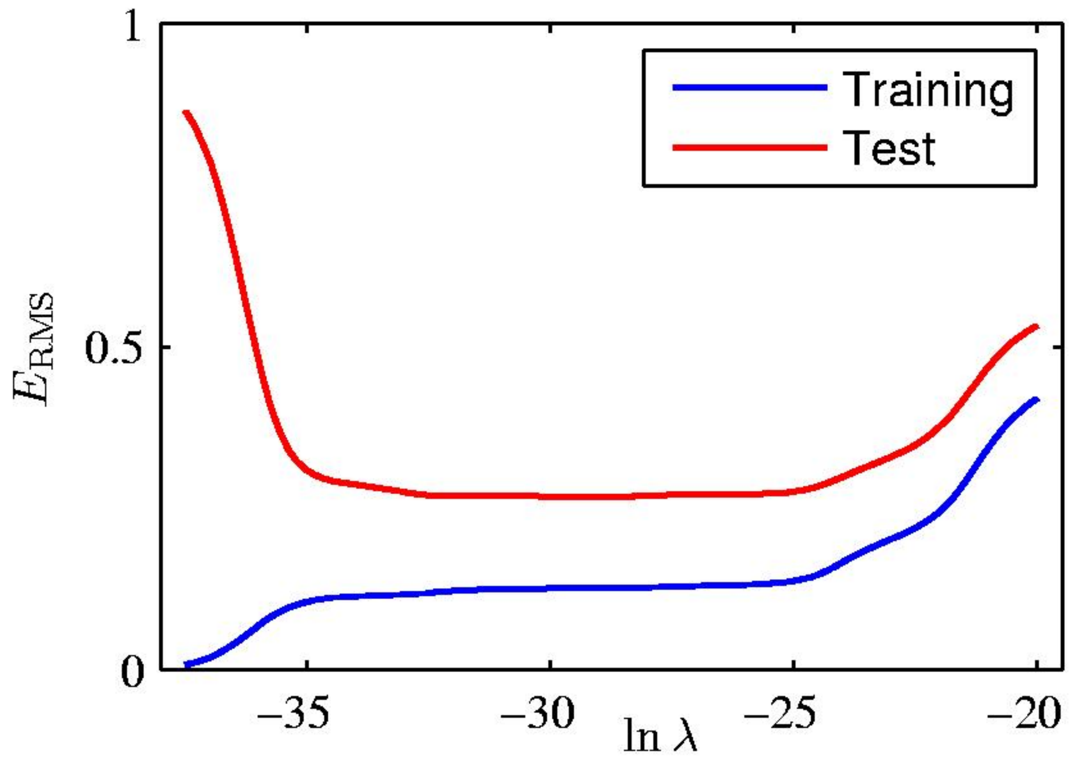
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$



Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	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

Binary Classification

Regression vs Classification

$y = f(x)$

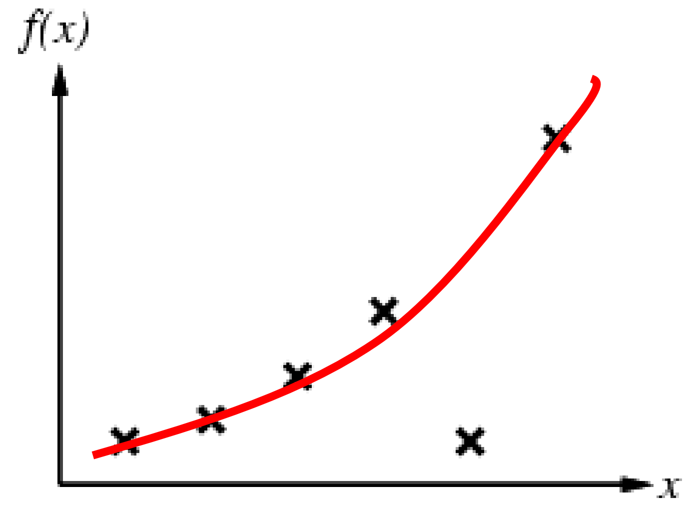
Regression:

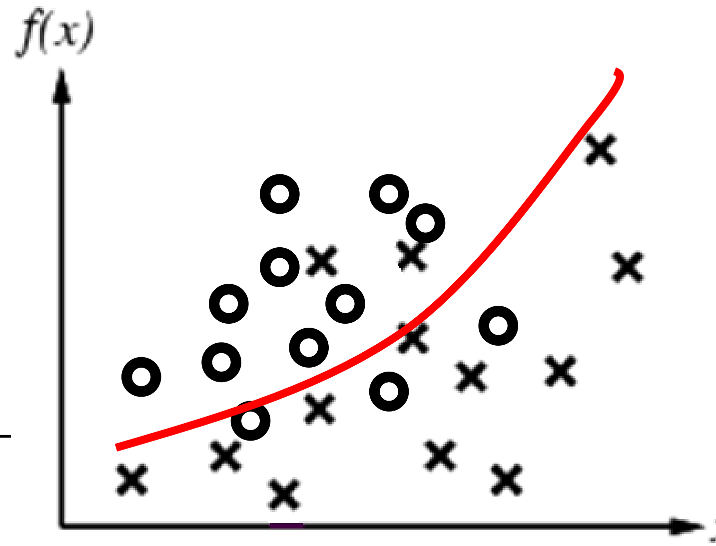
y is continuous

Classification:

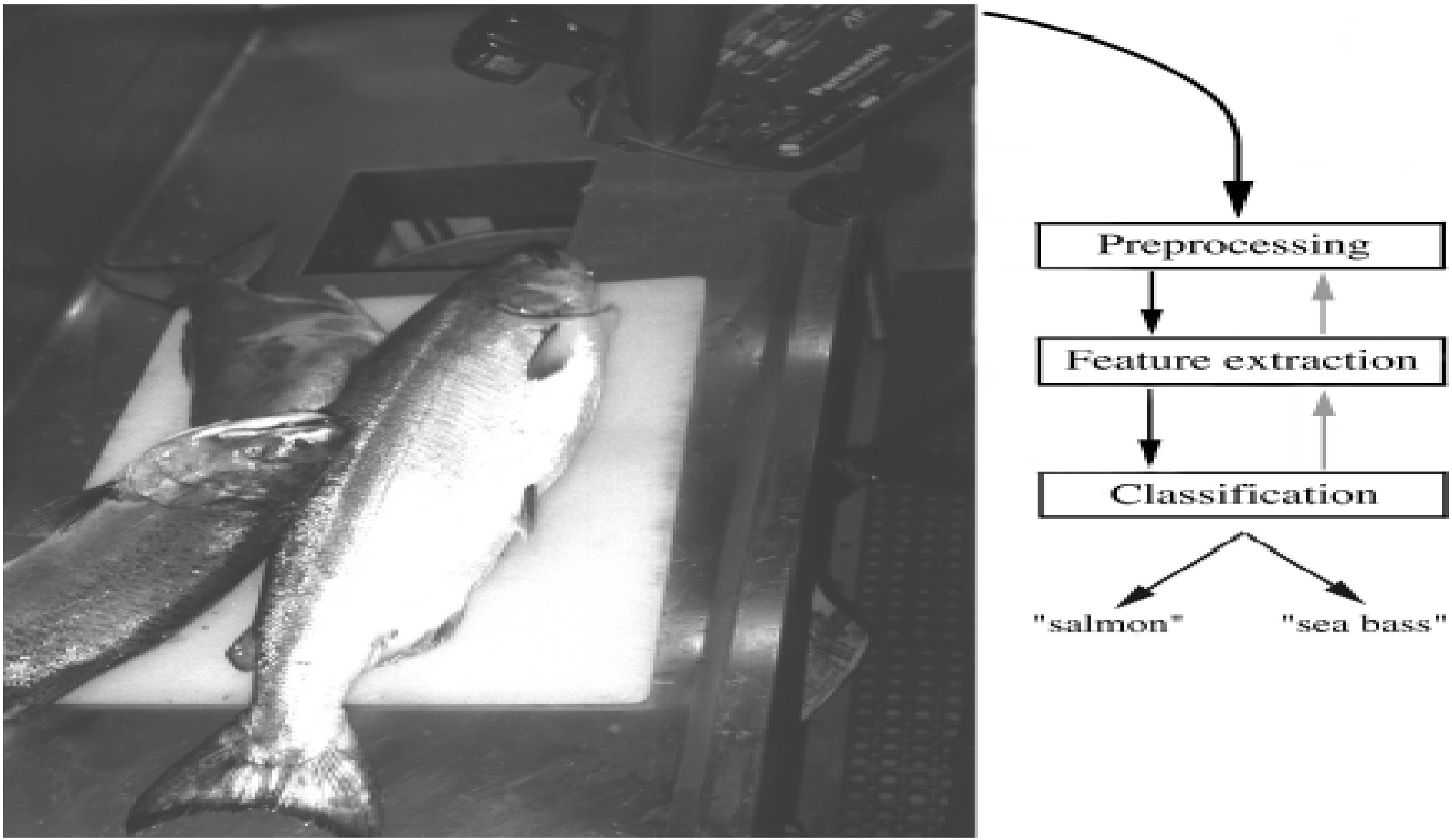
y : discrete values e.g. 0,1,2...
for classes C_0, C_1, C_2...

Binary Classification: two classes
$y \in \{0,1\}$

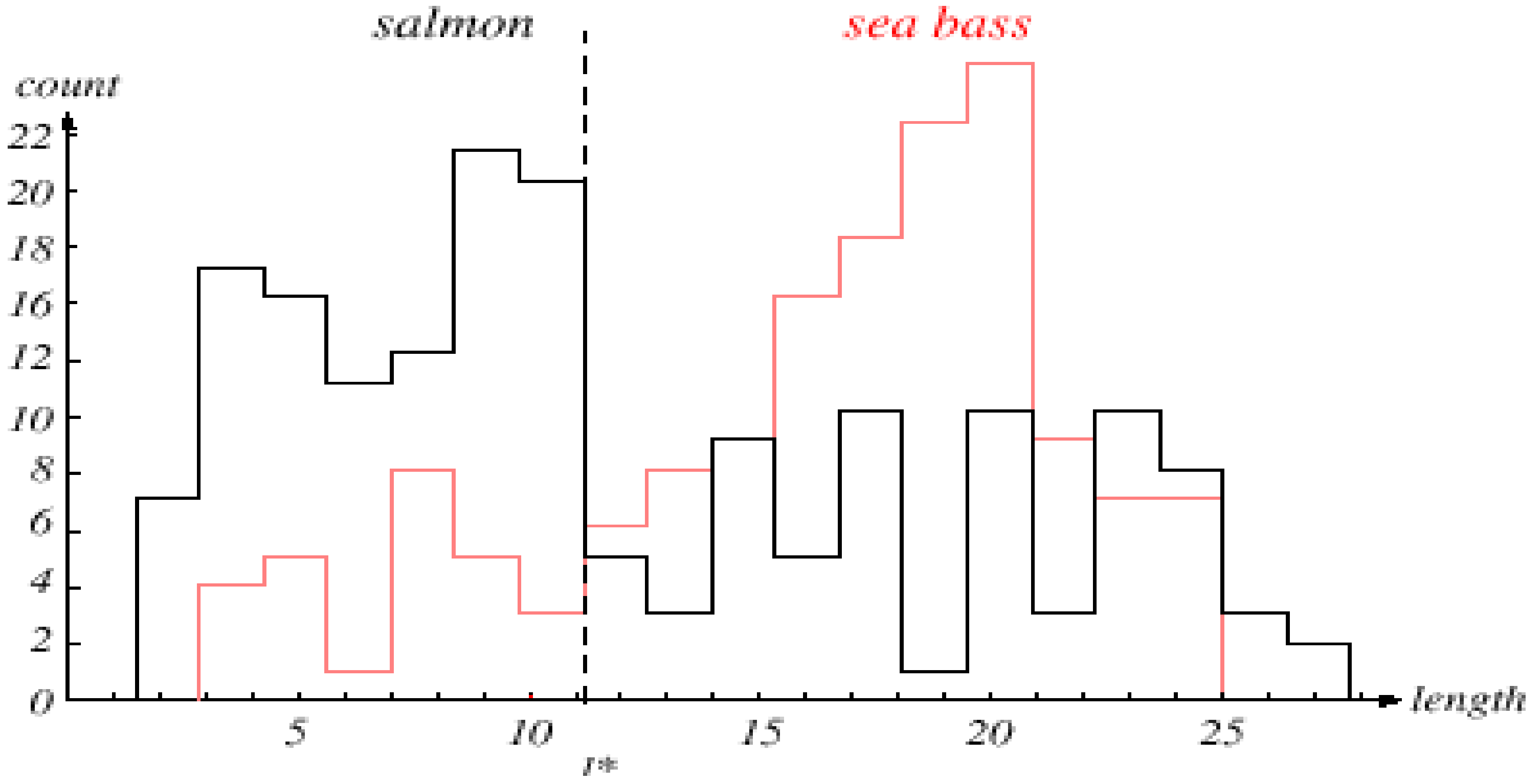




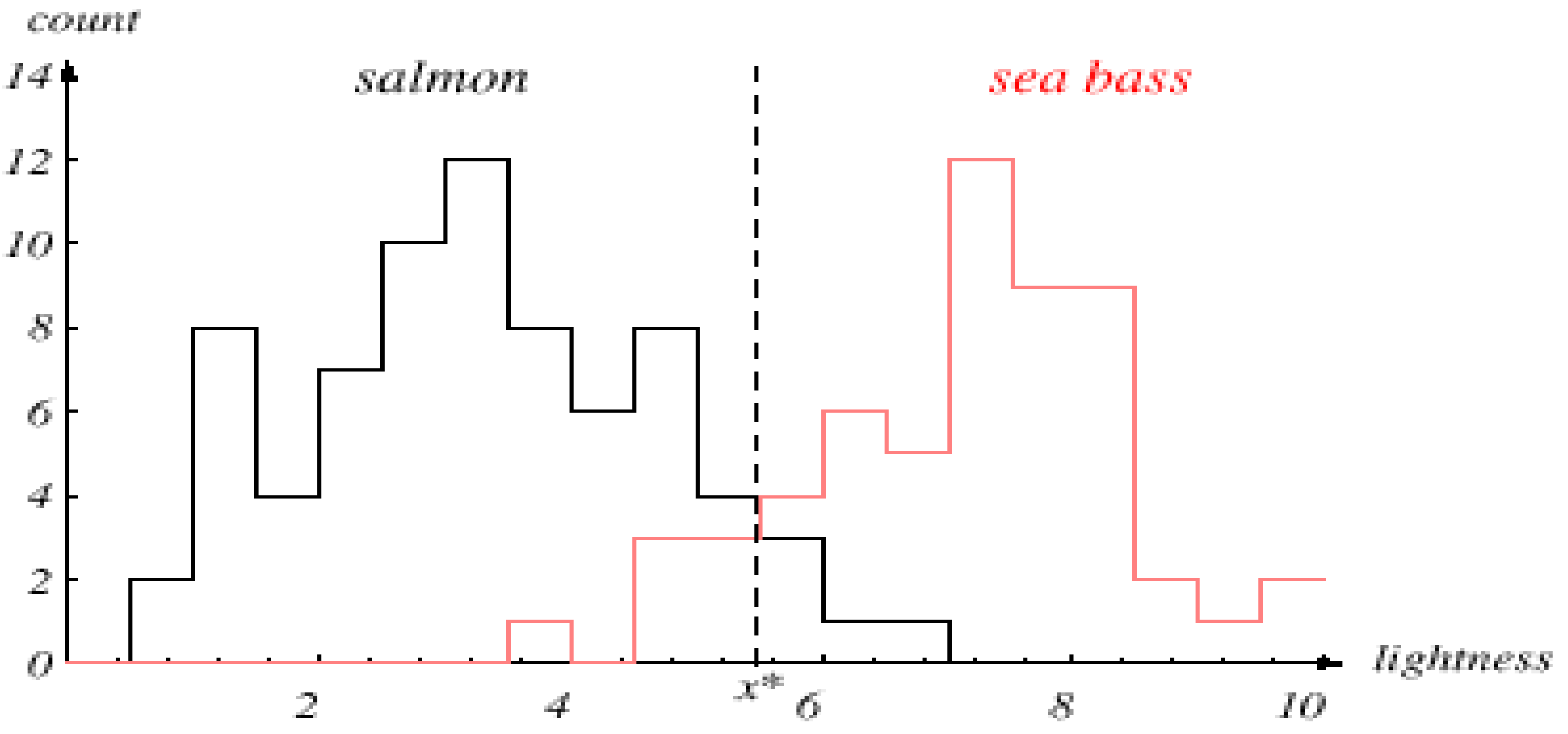
Binary Classification



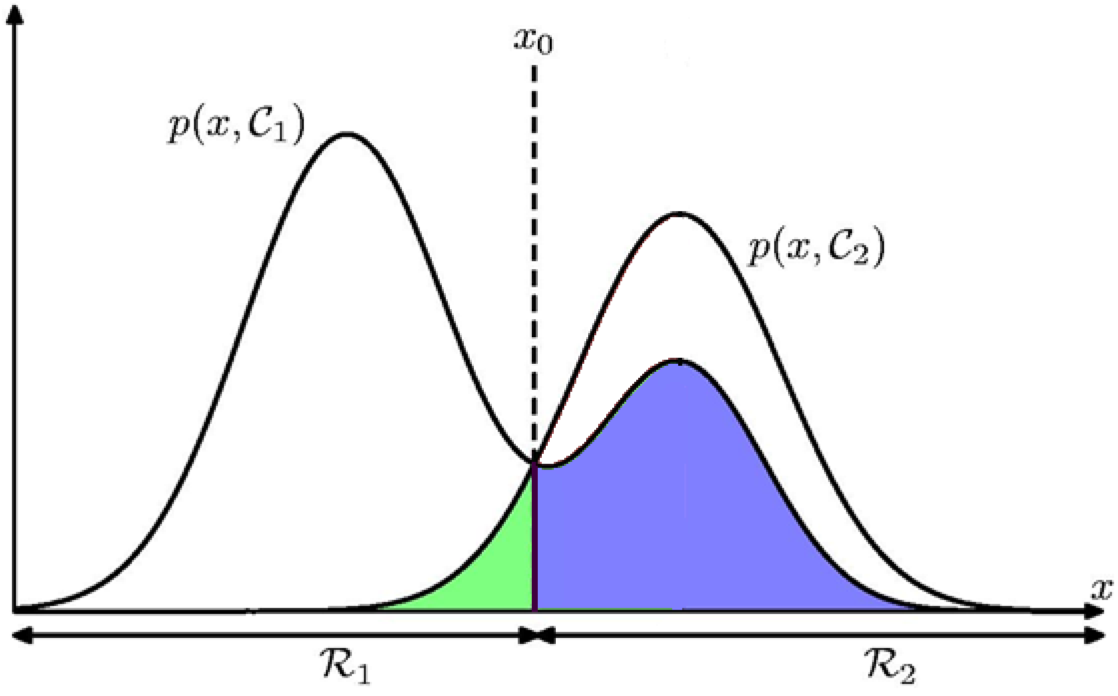
Feature : Length



Feature : Lightness



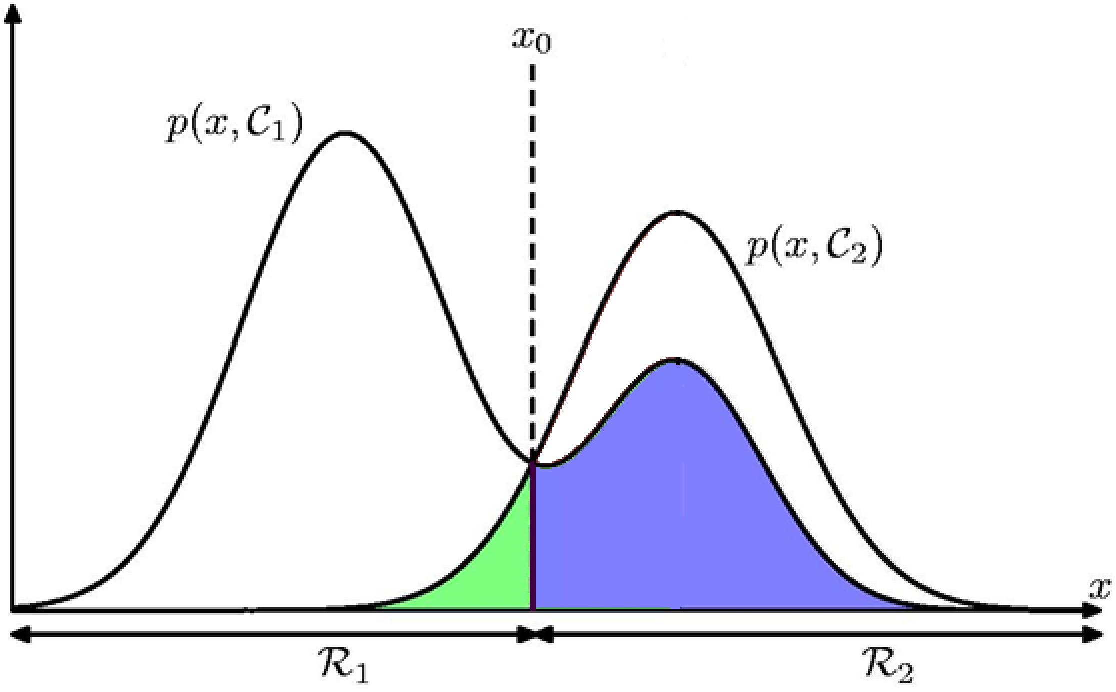
Minimize Misclassification



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

Precision / Recall

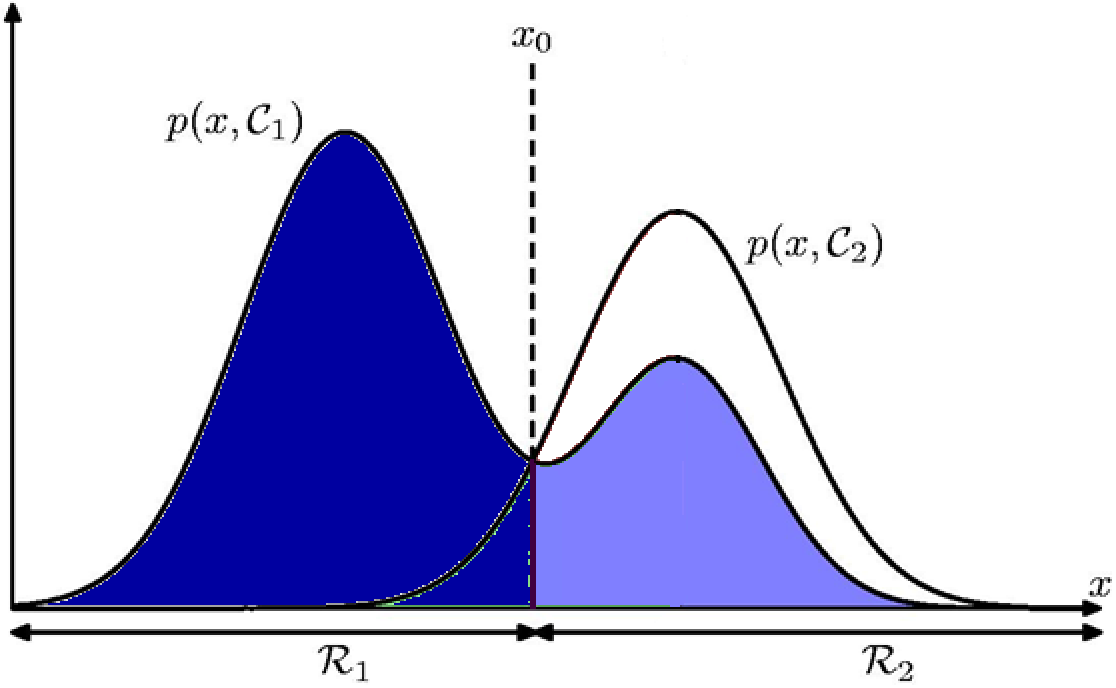
C1 : class of interest



Which is higher: Precision, or Recall?

Precision / Recall

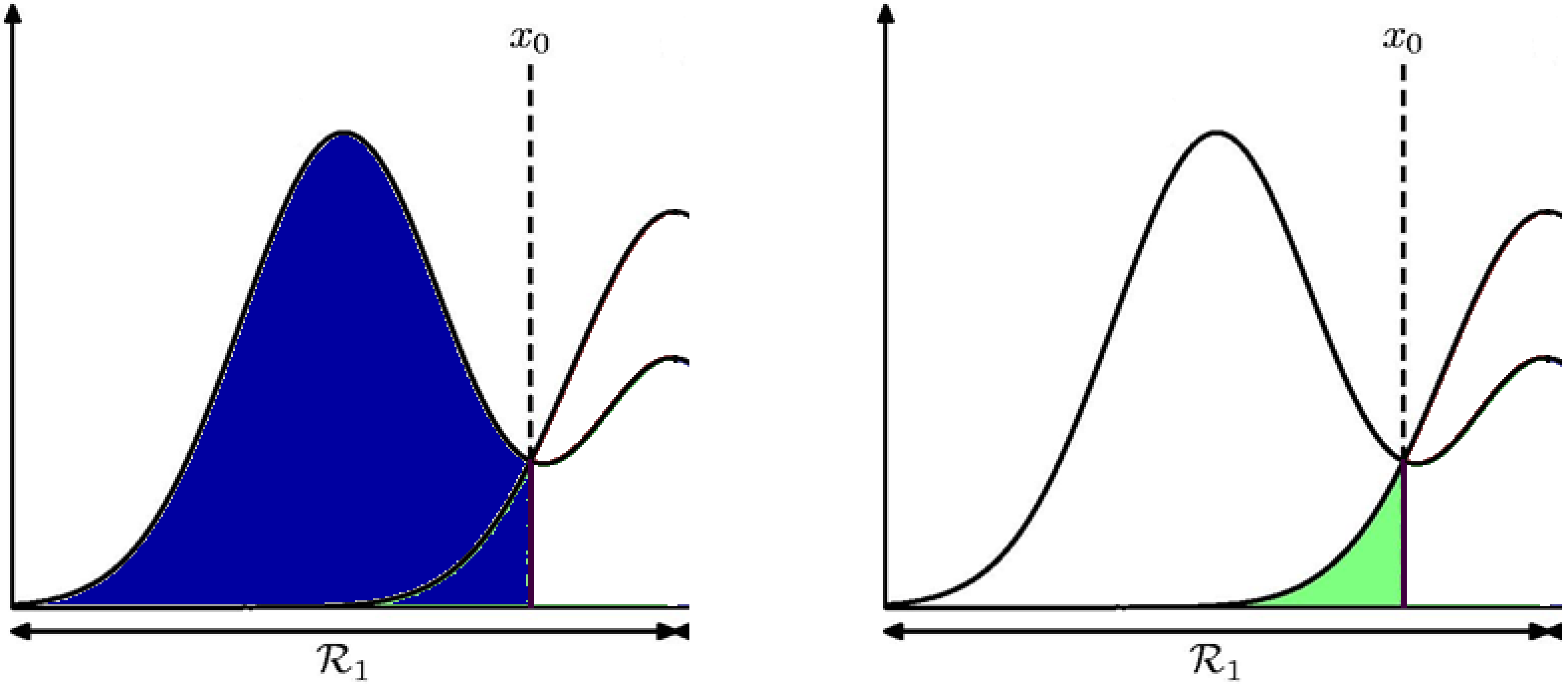
C1 : class of interest
(Positives)



Recall = TP / TP +FP

Precision / Recall

C1 : class of interest



Precision = TP / TP +FN

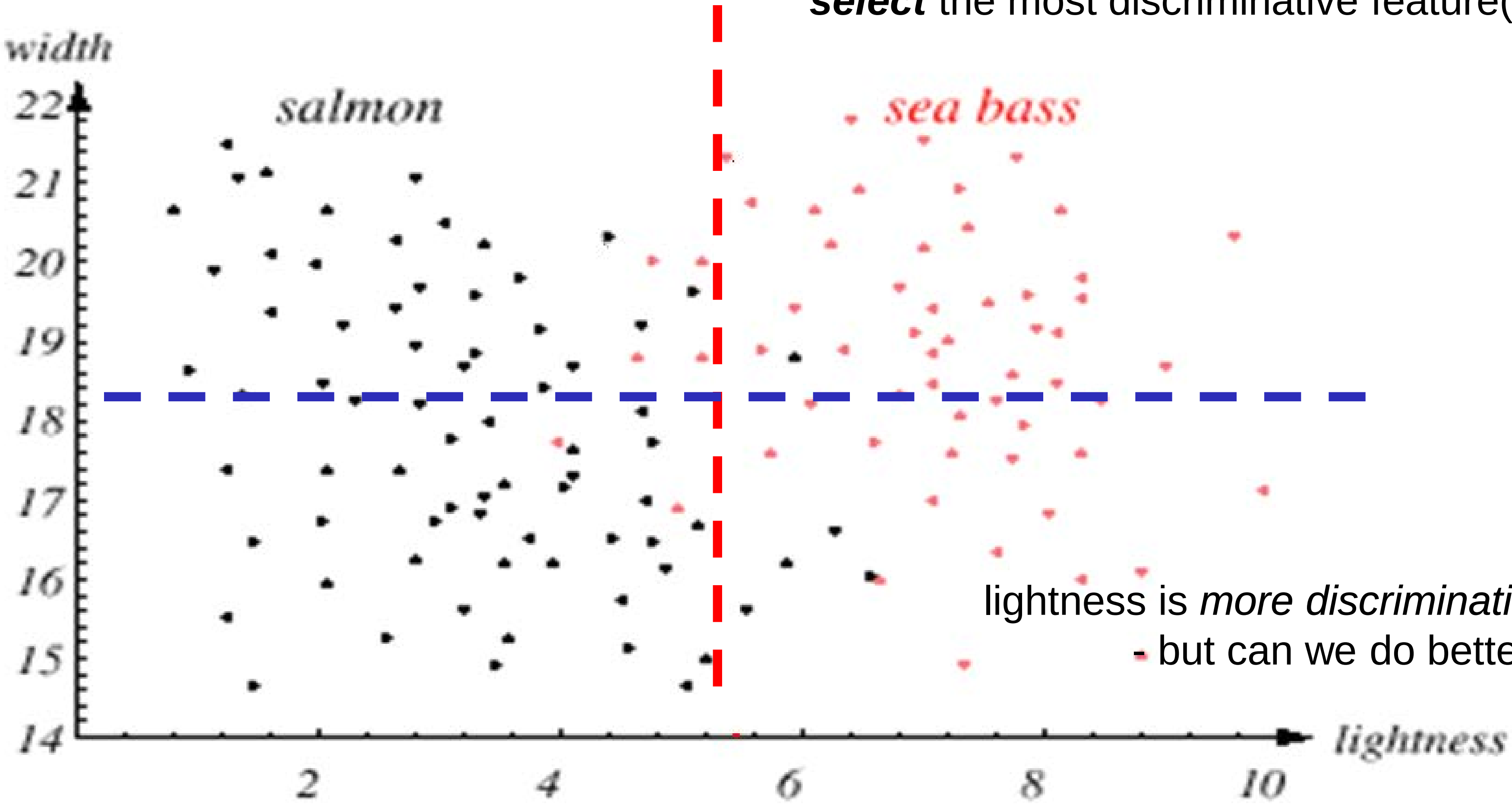
Decisions - Feature Space

- **Feature selection : which feature is maximally discriminative?**
  - Axis-oriented decision boundaries in feature space
  - Length – or – Width – or Lightness?

- **Feature Discovery: construct g(), defined on the feature space, for better discrimination**

Feature Selection: *width / lightness*

select the most discriminative feature(s)



lightness is *more discriminative*

- but can we do better?

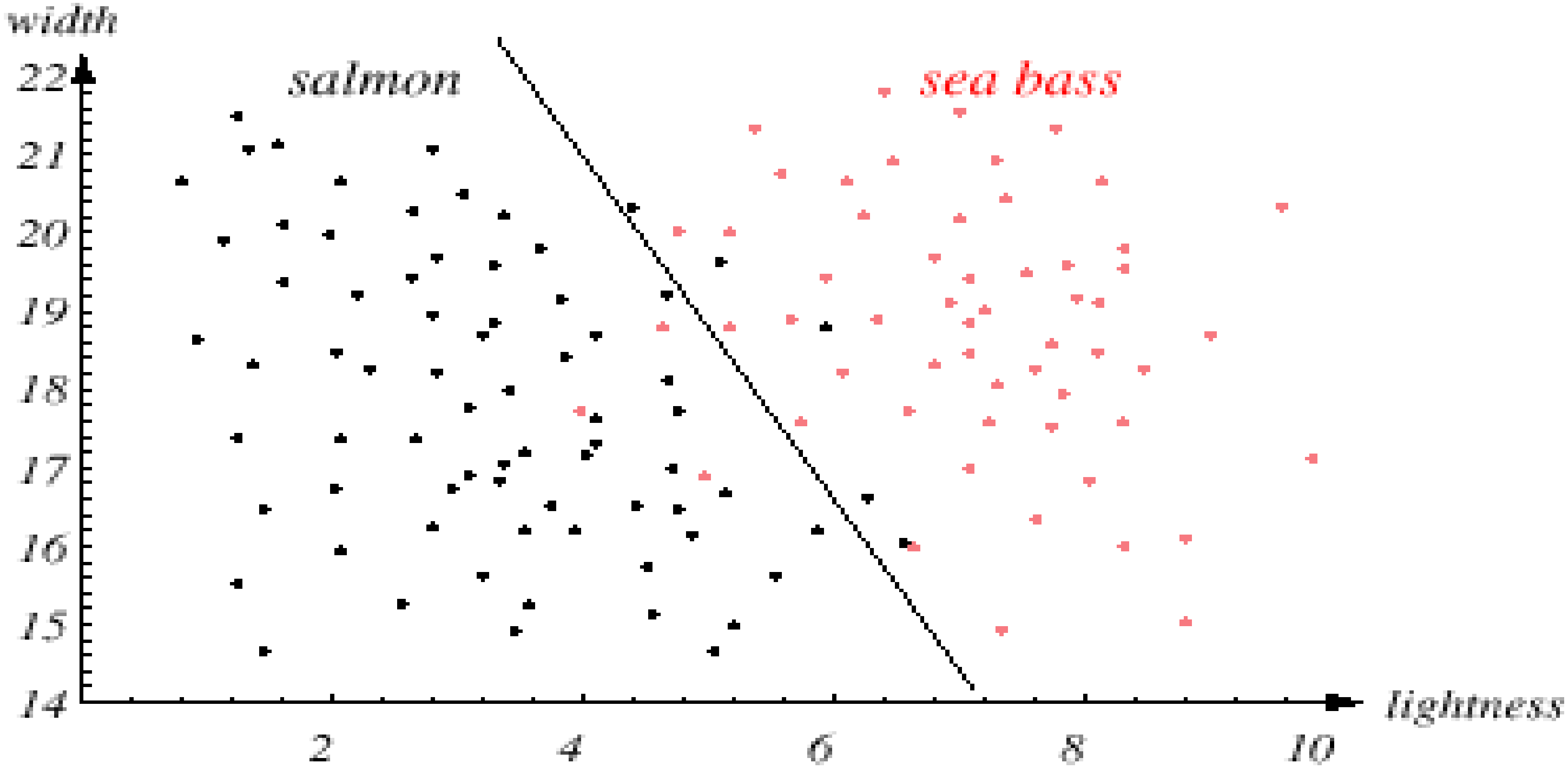
Feature Selection

- Feature selection : which feature is maximally discriminative?

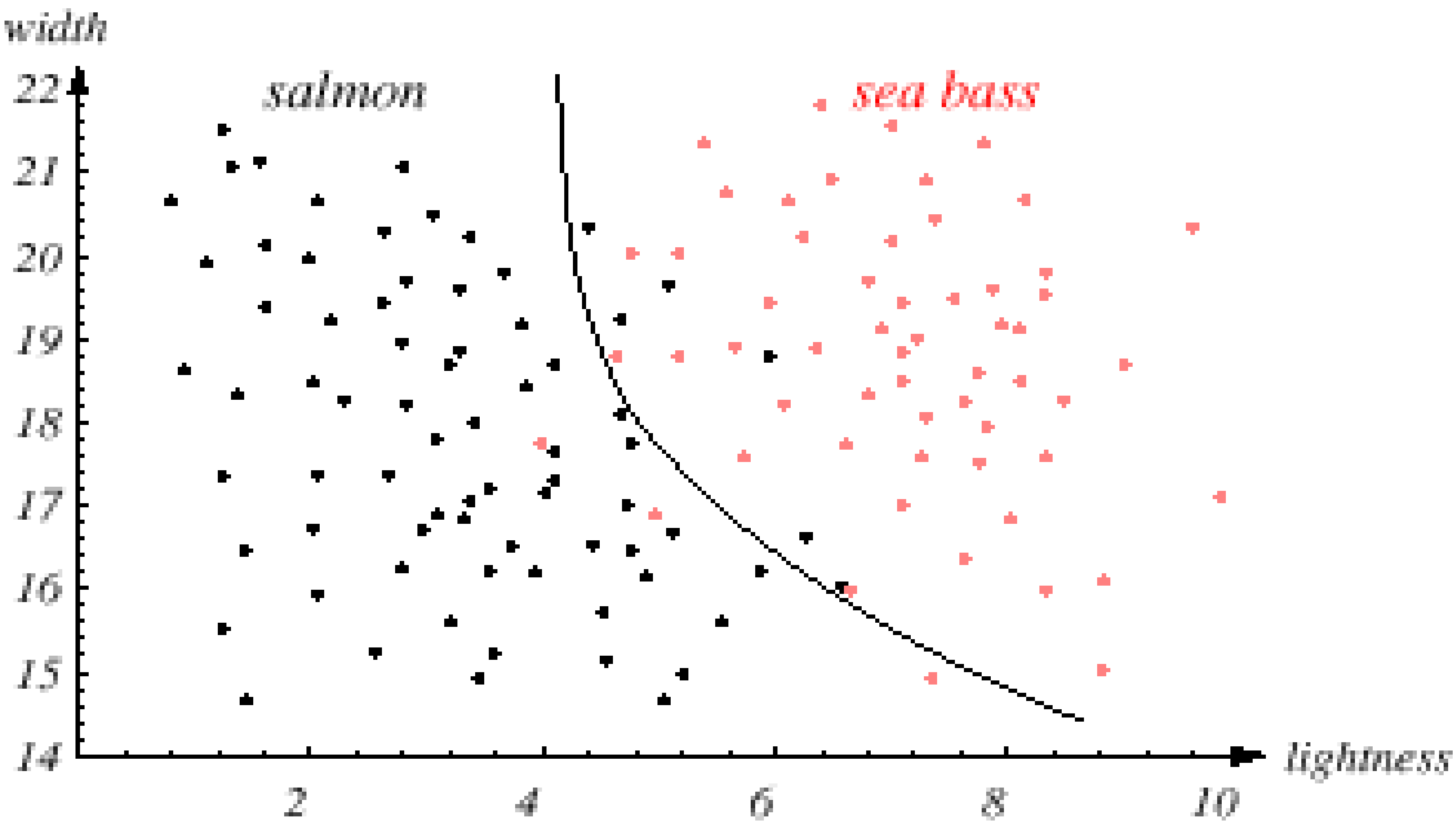
- Axis-oriented decision boundaries in feature space
- Length – or – Width – or Lightness?

- **Feature Discovery: discover discriminative function on feature space : g()**
  - combine aspects of length, width, lightness

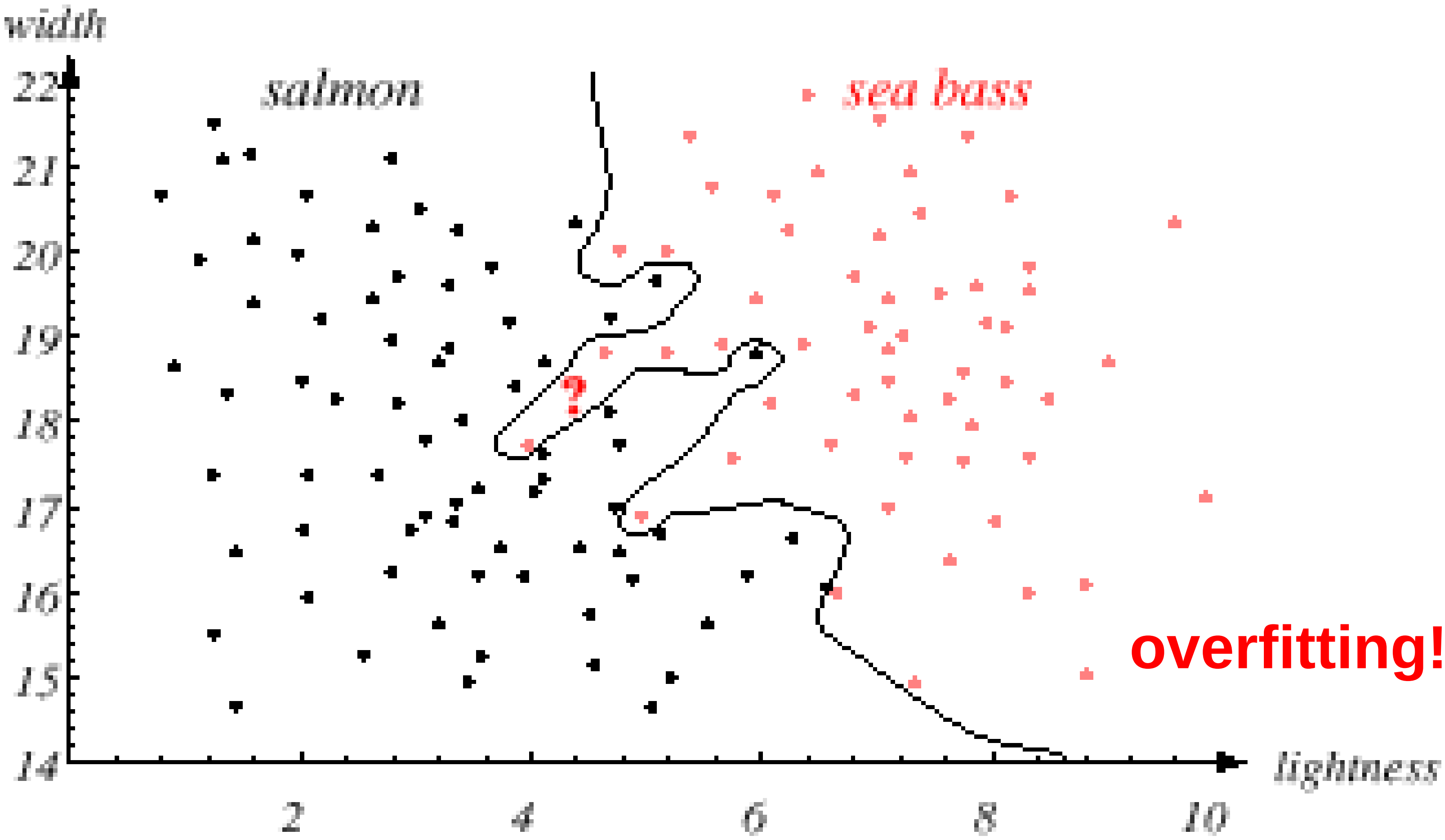
Feature Discovery : Linear



Decision Surface: non-linear



Decision Surface : non-linear



overfitting!

Learning process

- Feature set : representative? complete?

- Sample size : training set vs test set

- Model selection:

- Unseen data → overfitting?
- Quality vs Complexity
- Computation vs Performance

Best Feature set?

- Is it possible to describe the variation in the data in terms of a compact set of Features?

- **Minimum Description Length**

Probability Theory

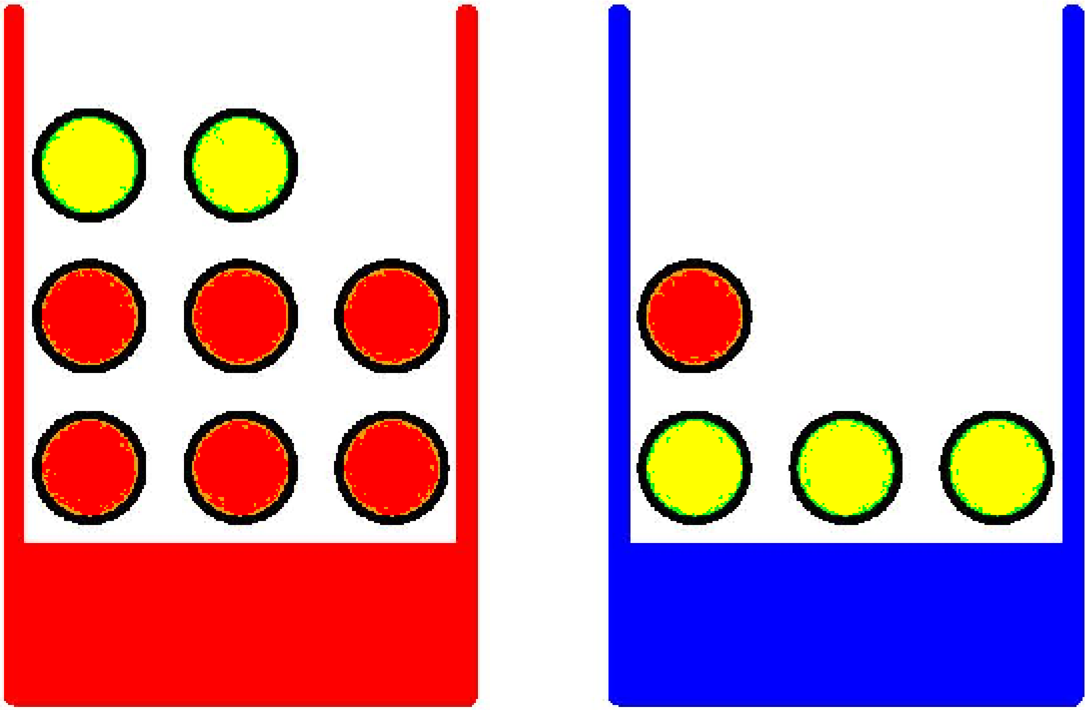
Learning = discovering regularities

- **Regularity** : repeated experiments:
  outcome not be fully predictable

  outcome = “possible world”
  set of all possible worlds = Ω

Probability Theory

Apples and Oranges



Sample Space

Sample ω = Pick two fruits,

e.g. Apple, then Orange

Sample Space $\Omega = \{(A,A), (A,O), (O,A),(O,O)\}$

= all possible worlds

Event e = set of possible worlds, $e \subseteq \Omega$

- e.g. second one picked is an apple

Learning = discovering regularities

- **Regularity** : repeated experiments:
  outcome not be fully predictable

- **Probability** p(e) : "the fraction of possible worlds
  in which e is true” i.e. outcome is event e

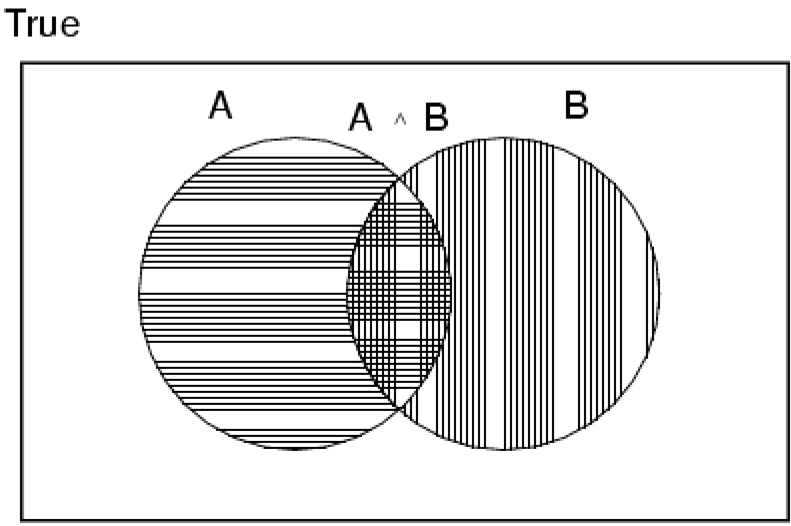
- **Frequentist** view : p(e) = limit as N $\rightarrow \infty$
- **Belief** view: in wager : equivalent odds
  (1-p):p that outcome is in e, or vice versa

Axioms of Probability

- **non-negative** : $p(e) \geq 0$

- **unit sum** $p(\Omega) = 1$

  i.e. no outcomes outside s



- **additive** : if e1, e2 are disjoint events (no common outcome):

$$p(e1) + p(e2) = p(e1 \cup e2)$$

ALT:

$$p(e1 \vee e2) = p(e1) + p(e2) - p(e1 \wedge e2)$$

Why probability theory?

different methodologies attempted for uncertainty:

- Fuzzy logic
- Multi-valued logic
- Non-monotonic reasoning

But **unique property** of probability theory:

If you gamble using probabilities you have the best chance in a wager. [de Finetti 1931]

=> if opponent uses some other system, he's more likely to lose

Ramsay-diFinetti theorem (1931)

If agent X's degrees of belief are rational, then X 's degrees of belief function defined by fair betting rates is (formally) a probability function

Fair betting rates: opponent decides which side one bets on

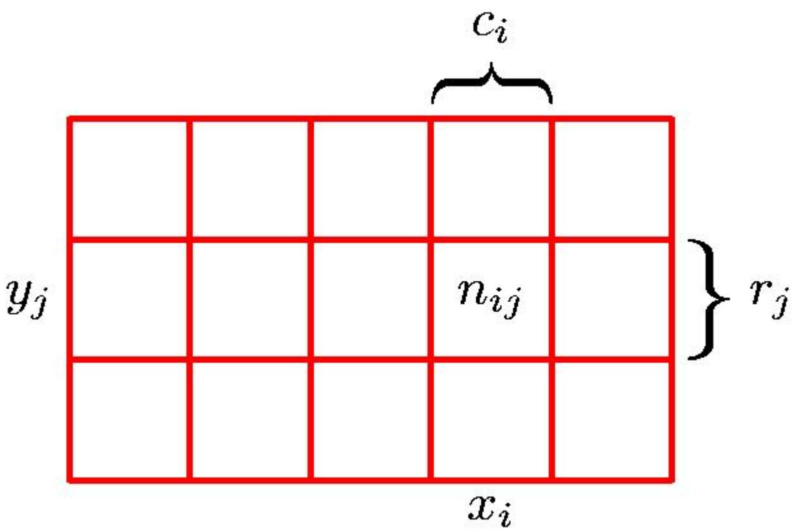
Proof: fair odds result in a function pr () that satisifies the Kolmogrov axioms:

Normality : pr(S) >=0

Certainty : pr(T)=1

Additivity : pr (S1 v S2 v..)= Σ(Si)

Joint vs. conditional probability



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

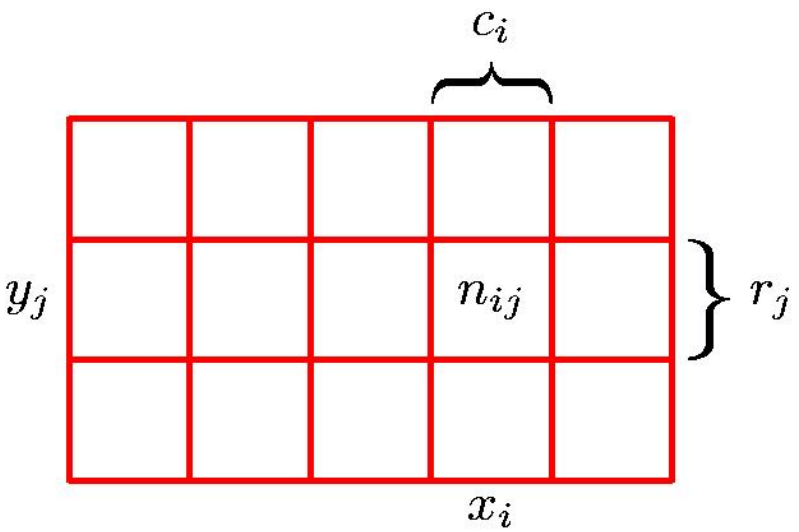
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

Rules of Probability

Sum Rule $$p(X) = \sum_Y p(X, Y)$$

Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Example

A disease *d* occurs in 0.05% of population. A test is 99% effective in detecting the disease, but 5% of the cases test positive in absence of *d.*

10000 people are tested. How many are expected to test positive?

$p(d) = 0.0005$; $p(t/d) = 0.99$; $p(t/\sim d) = 0.05$

$p(t) = p(t,d) + p(t,\sim d)$ [Sum Rule]

$= p(t/d)p(d) + p(t/\sim d)p(\sim d)$ [Product Rule]

$= 0.99 * 0.0005 + 0.05 * 0.9995 = 0.0505$ → **505** +ve

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior

Bayes’ Theorem

Thomas Bayes (c.1750):

how can we infer causes from effects?

How can one learn the probability of a future event if one knew only

how many times it had (or had not) occurred in the past?

as new evidence comes in --> prob knowledge improves.

e.g. throw a die. guess is poor (1/6)

throw die again. is it > or < than prev? Can improve guess.

throw die repeatedly. can improve prob of guess quite a lot.

Hence: initial estimate (*prior* belief *P(h)*, not well formulated)

+ new evidence (support) – compute likelihood *P(data|h)*

→ improved estimate (*posterior P(h|data)*)

Example

A disease *d* occurs in 0.05% of population. A test is 99% effective in detecting the disease, but 5% of the cases test positive in absence of *d.*

If you are tested +ve, what is the probability you have the disease?

p(d/t) = p(d) . p(t/d) / p(t) ; p(t) = 0.0505

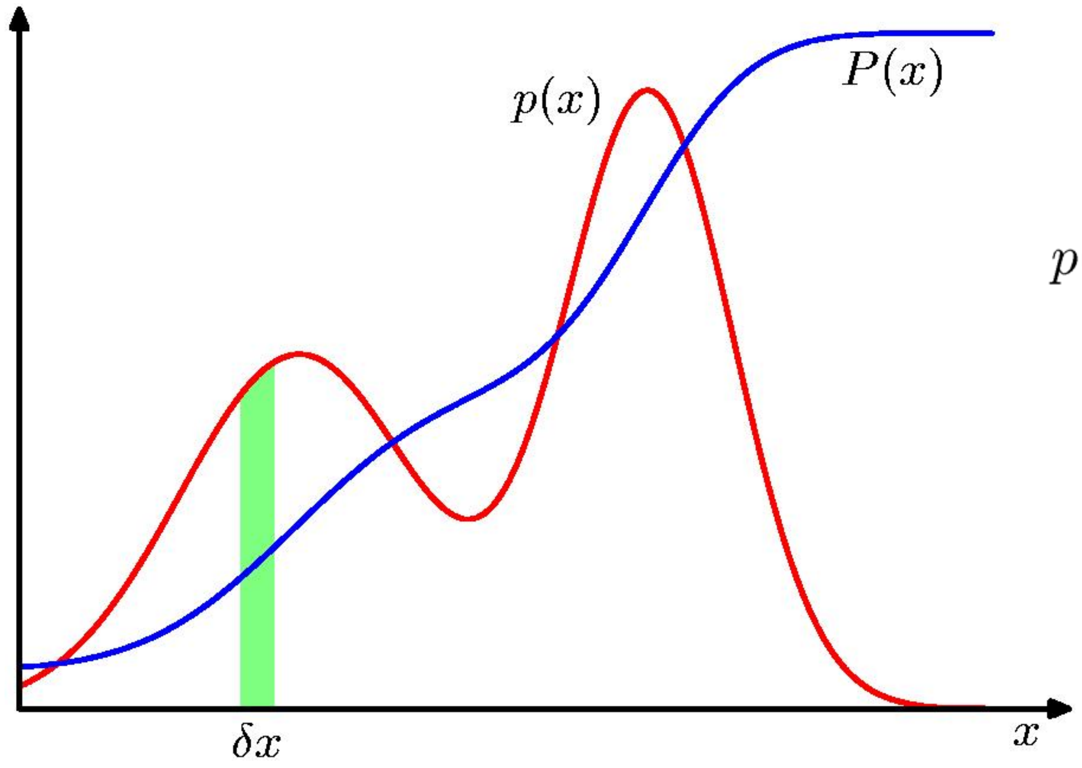
p(d/t) = 0.0005 * 0.99 / 0.0505 = 0.0098 (about **1%**)

if 10K people take the test, E(d) = 5

FPs = 0.05 * 9995 = 500

TPs = 0.99 * 5 = 5. → only **5/505** have *d*

Probability Densities



$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

discrete x

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

continuous x

Frequentist approximation w unbiased sample

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

(both discrete / continuous)

Variances and Covariances

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$\mathbb{E}_x[f(x,y)]$: Sum over x p(x)f(x,y) --> is a function of y

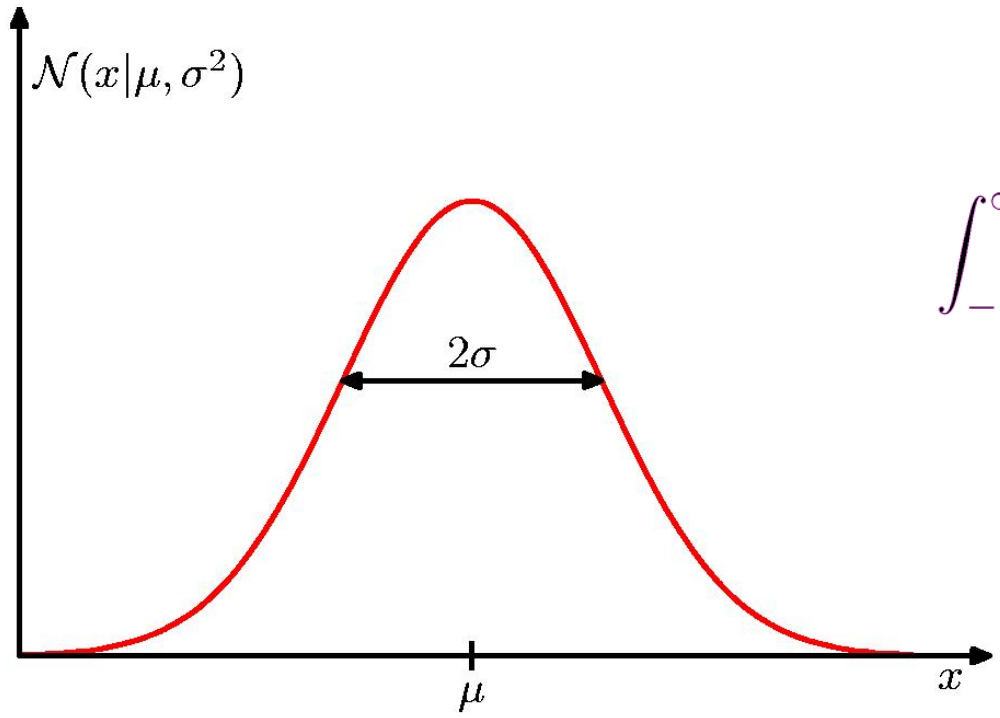
$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

Gaussian Distribution

The Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

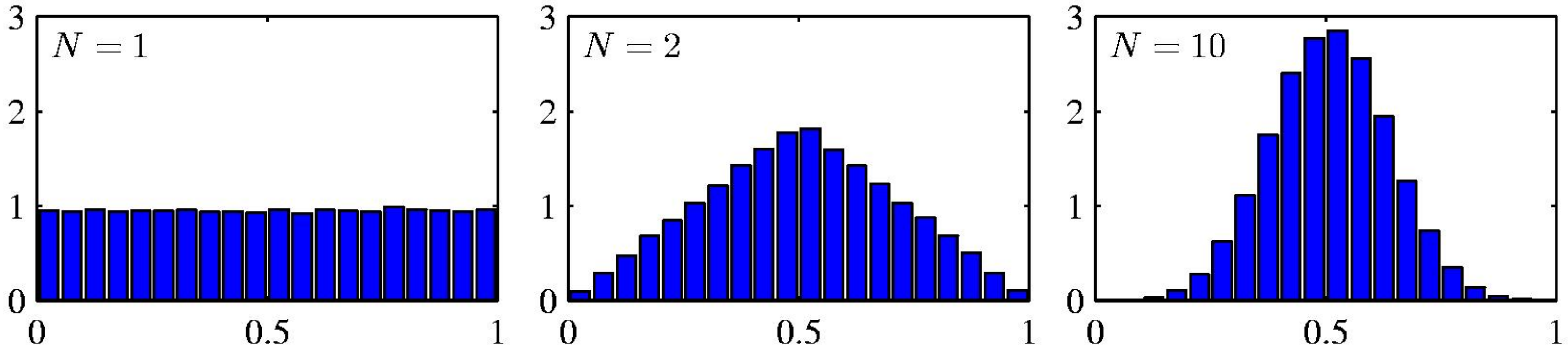
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Central Limit Theorem

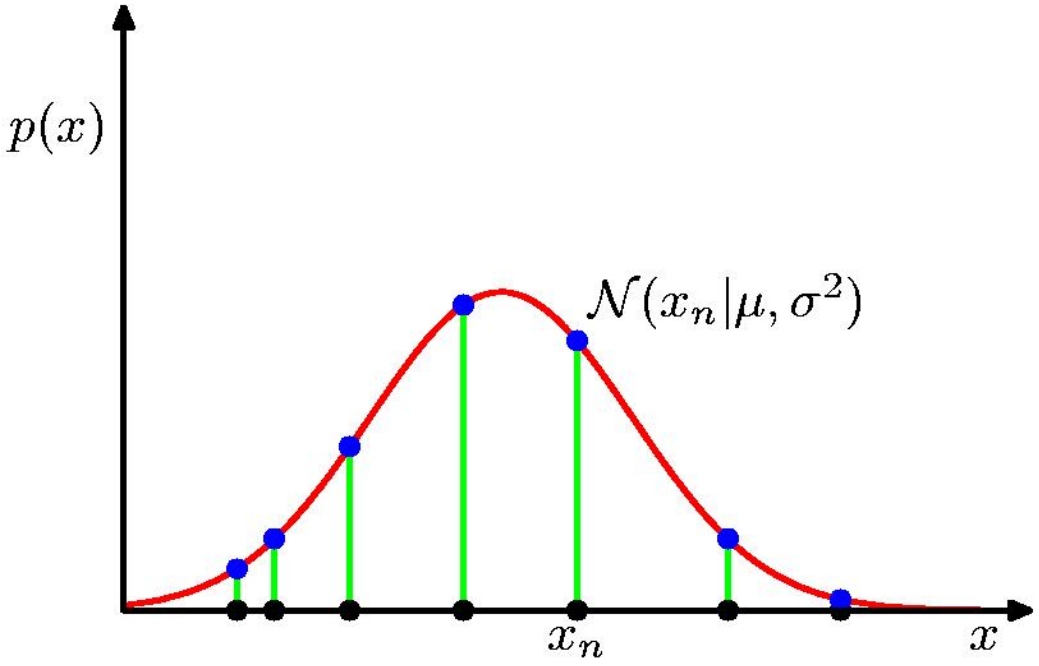
Distribution of sum of N i.i.d. random variables becomes increasingly Gaussian for larger N.

Example: N uniform [0,1] random variables.



Gaussian Parameter Estimation

Observations assumed to be indpendently drawn from same distribution (i.i.d)



Likelihood function

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

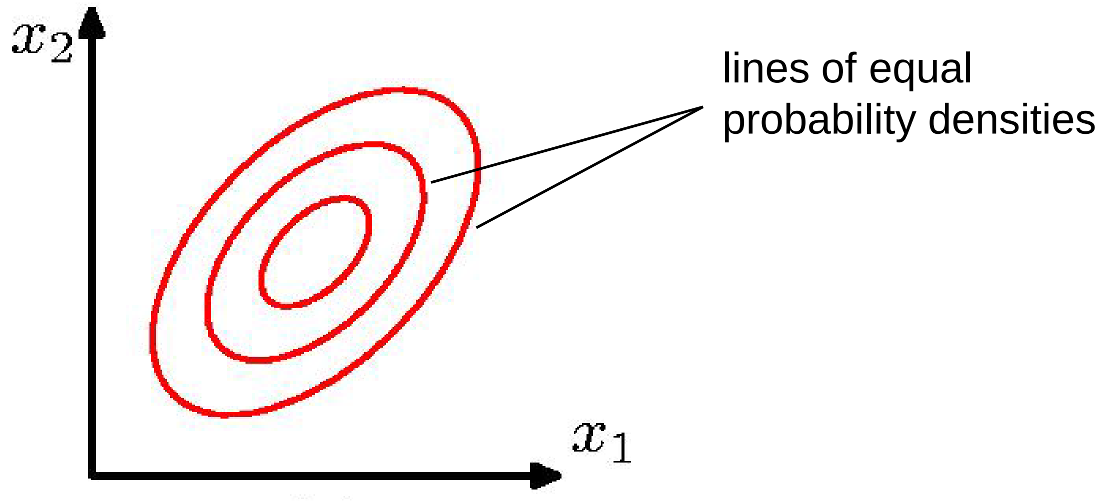
Maximum (Log) Likelihood

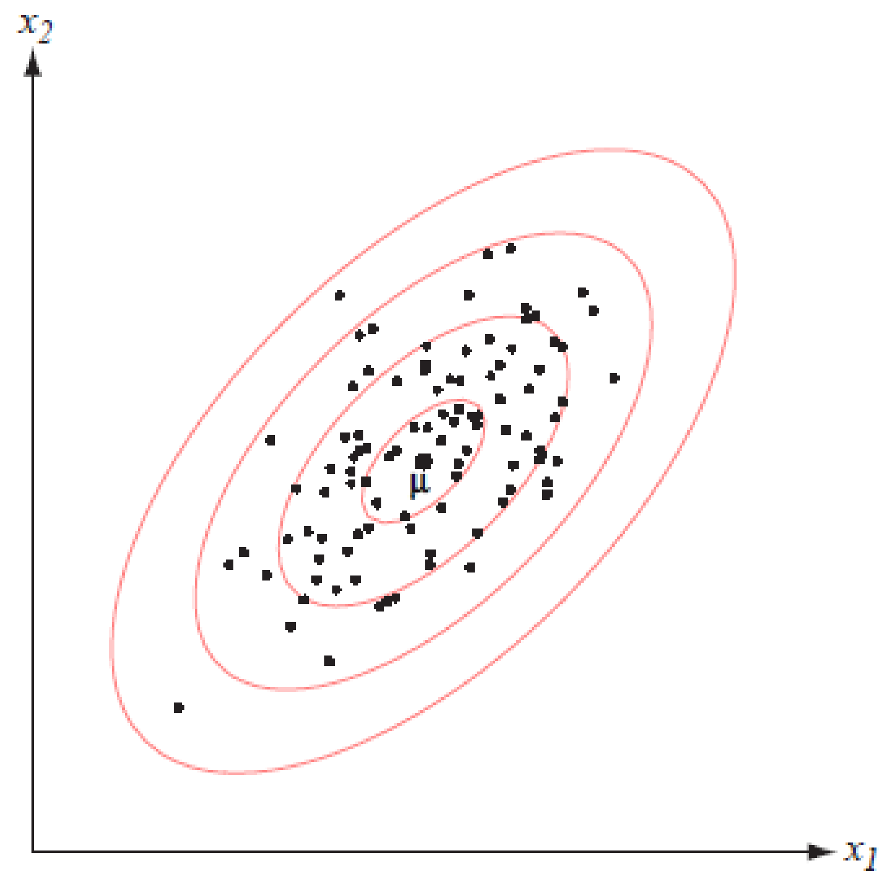
$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

The Multivariate Gaussian

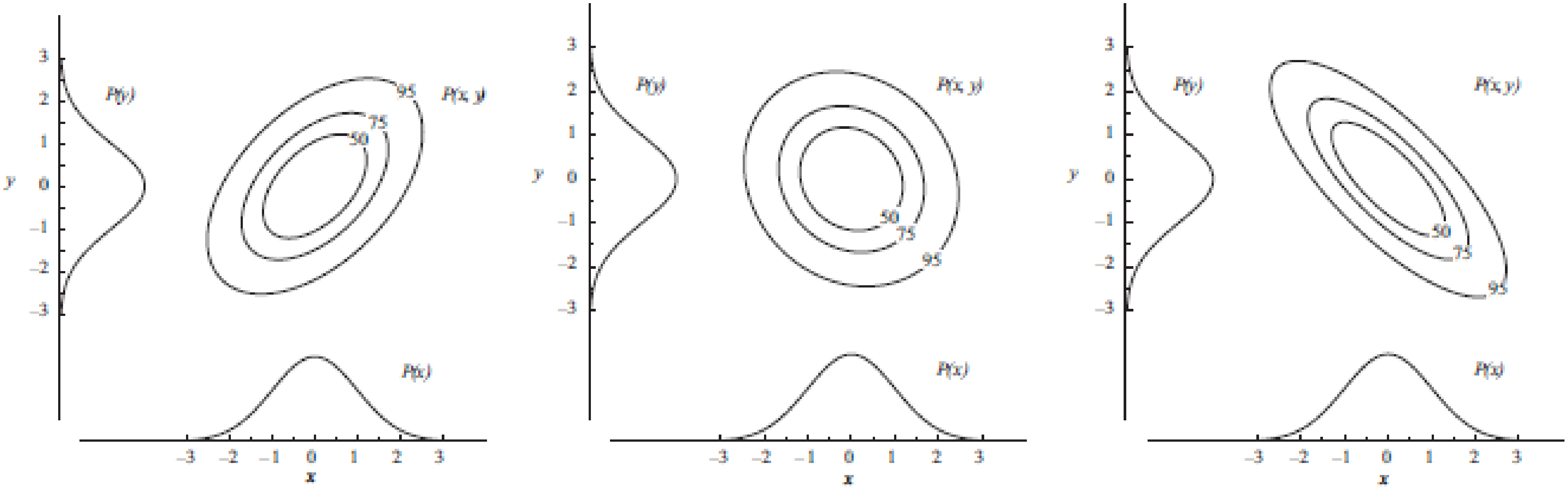
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$





x_2
μ
x_1

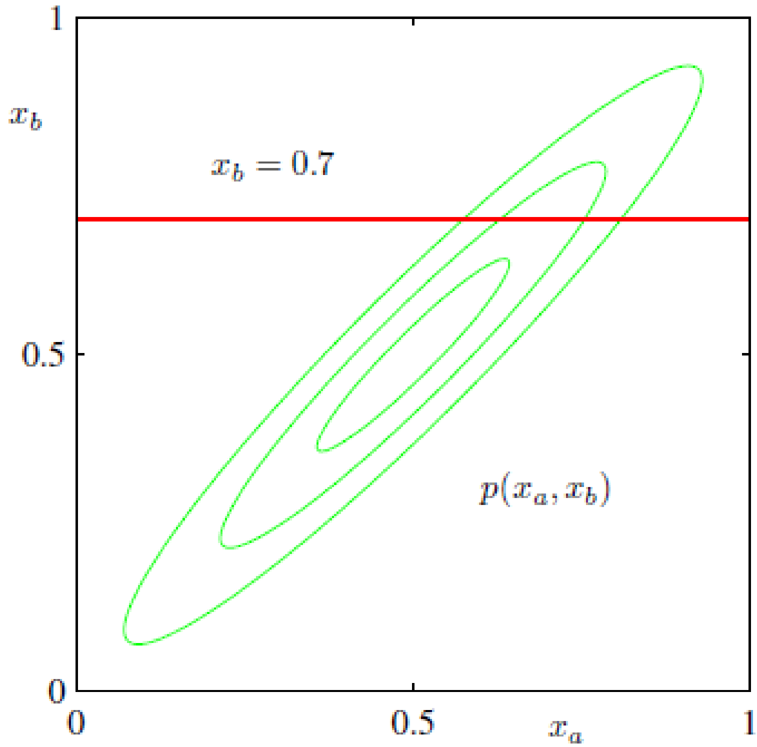
Multivariate distribution

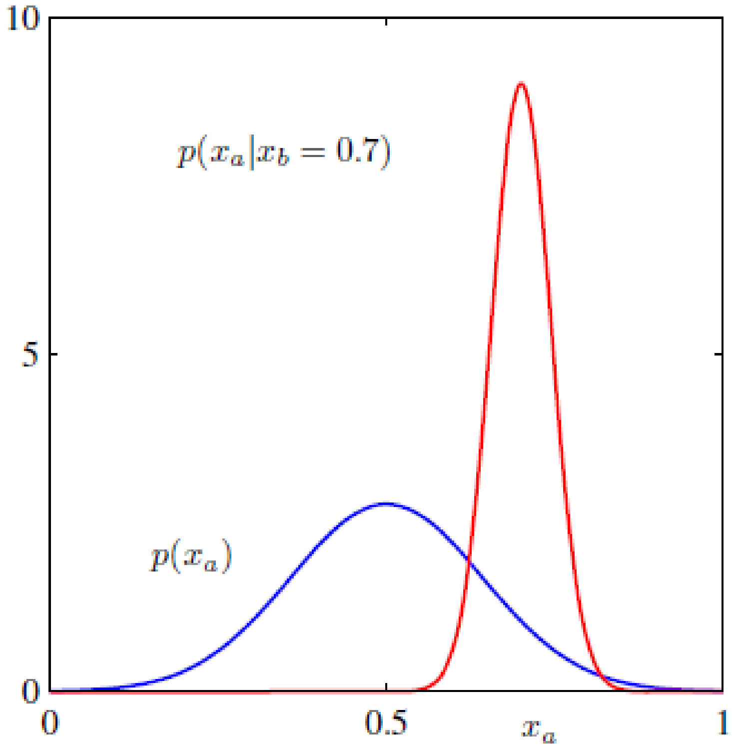


joint distribution P(x,y) varies considerably
though marginals P(x), P(y) are identical

estimating the joint distribution requires
much larger sample: $O(n^k)$ vs nk

Marginals and Conditionals

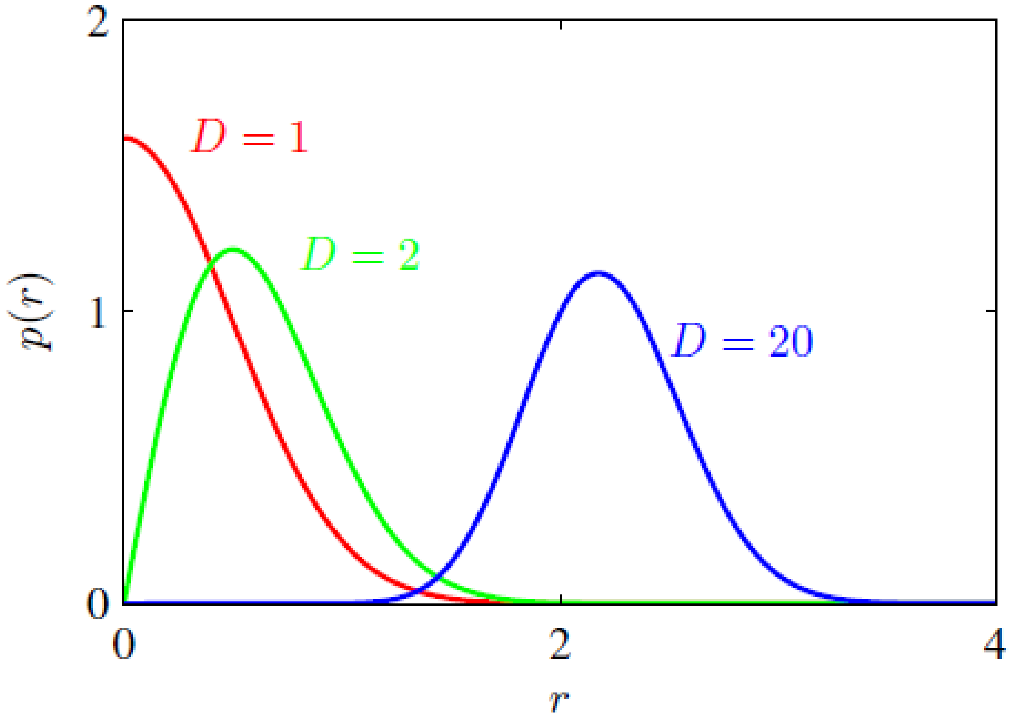




marginals P(x), P(y) are gaussian
conditional P(x|y) is also gaussian

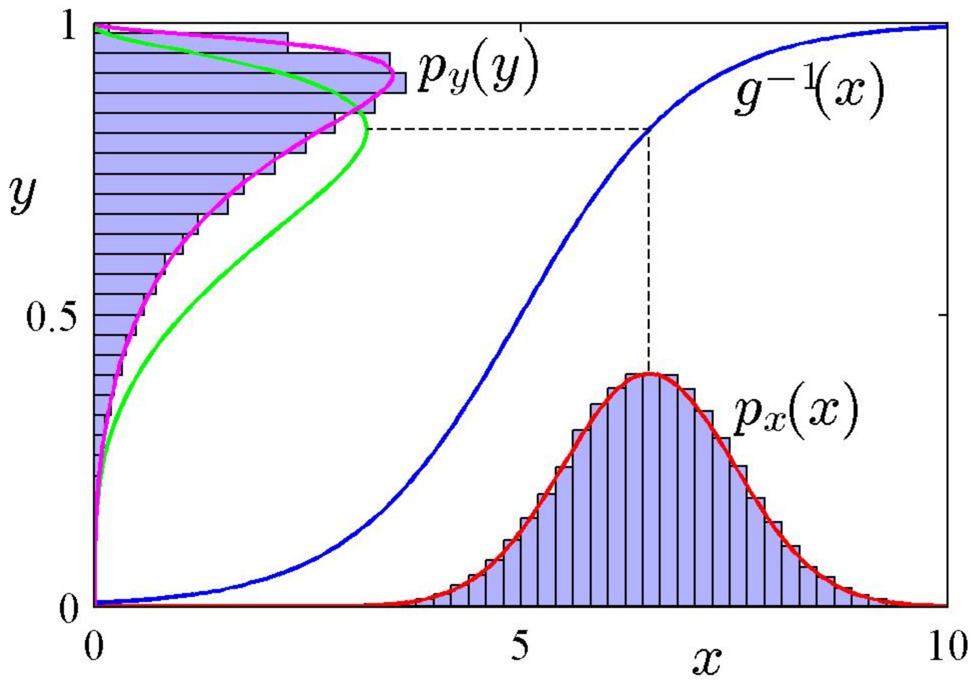
Non-intuitive in high dimensions

As dimensionality increases, bulk of data moves away from center



Gaussian in polar coordinates;
p(r)δr : prob. mass inside annulus *δr* at *r*.

Change of variable x=g(y)



$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$$
$$= p_x(g(y))\,|g'(y)|$$

Bernoulli Process

Successive Trials – e.g. Toss a coin three times:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

Probability of k Heads:

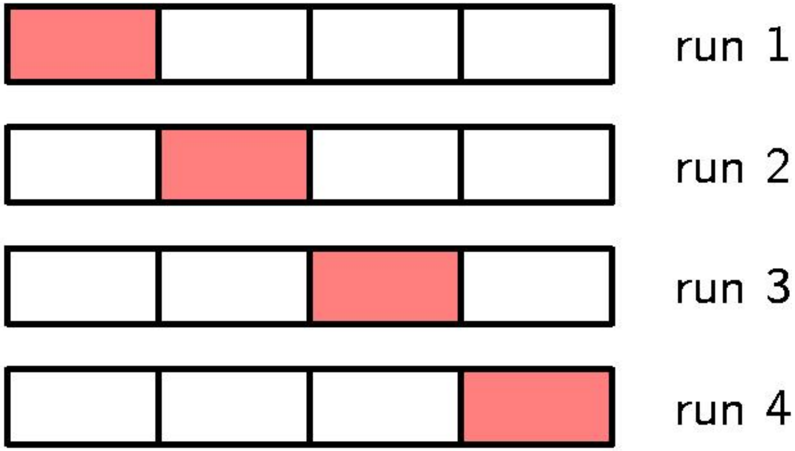
k	0	1	2	3
P(k)	1/8	3/8	3/8	1/8

$$P(k) = \binom{n}{k} p^k q^{n-k}$$

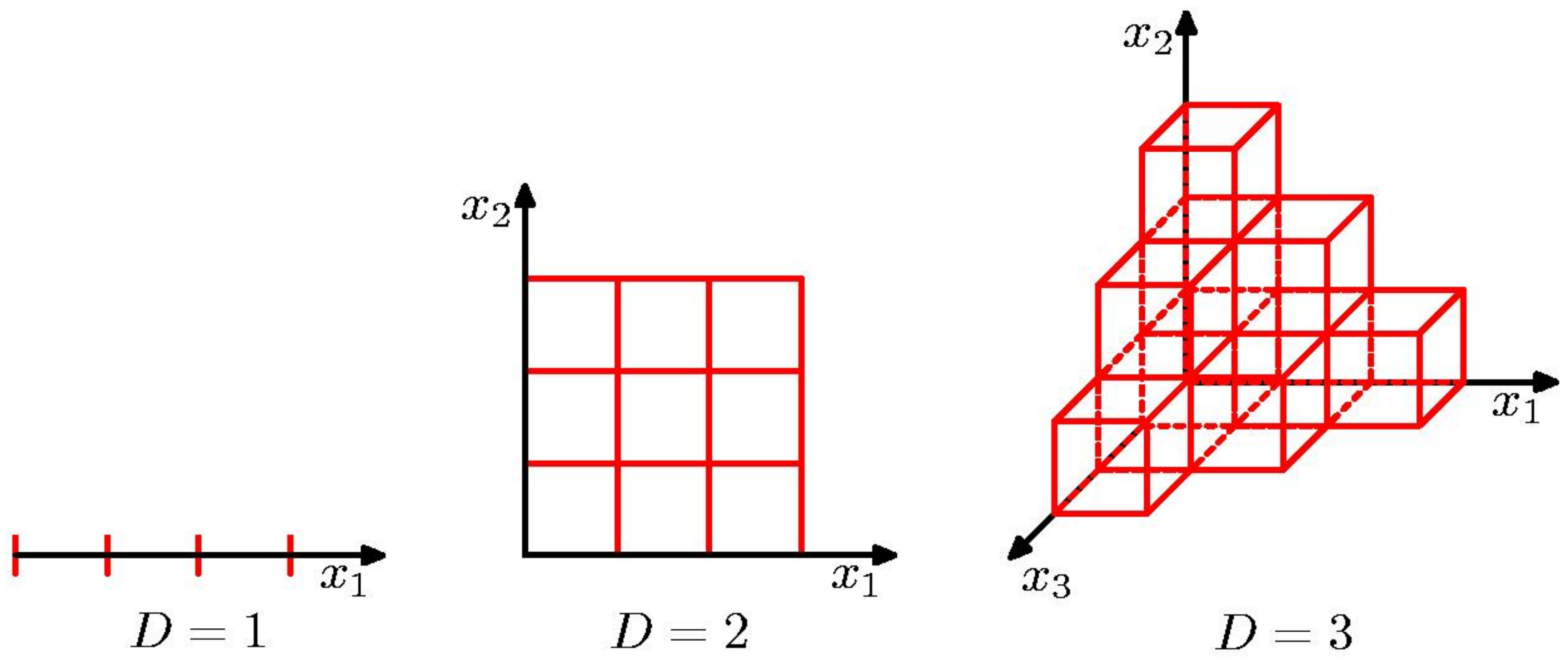
Model Selection

Model Selection

Cross-Validation



Curse of Dimensionality

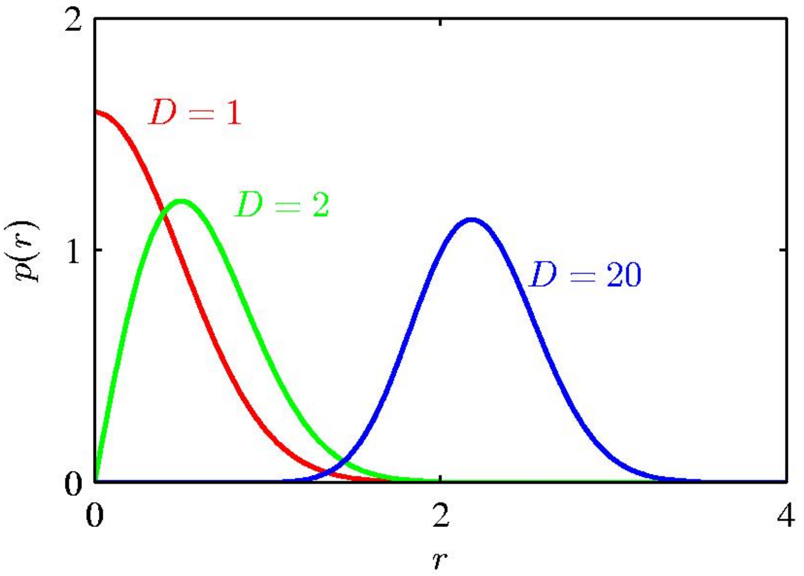


Curse of Dimensionality

Polynomial curve fitting, M = 3

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Gaussian Densities in higher dimensions

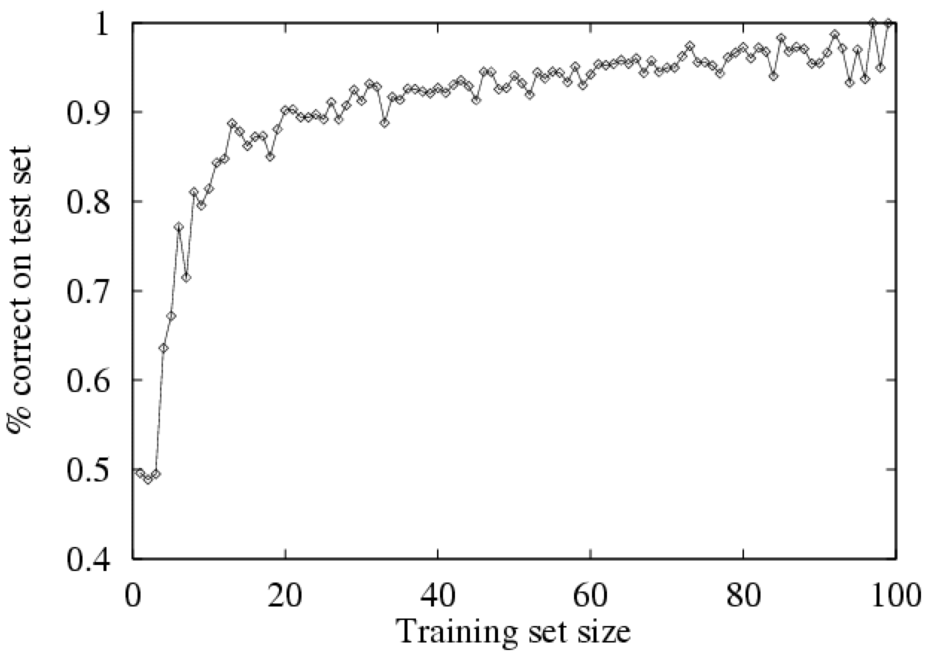


Performance measurement

- **How do we know that *h* ≈ *f* ?**
  1. Use theorems of computational/statistical learning theory
  2. Try *h* on a new test set of examples

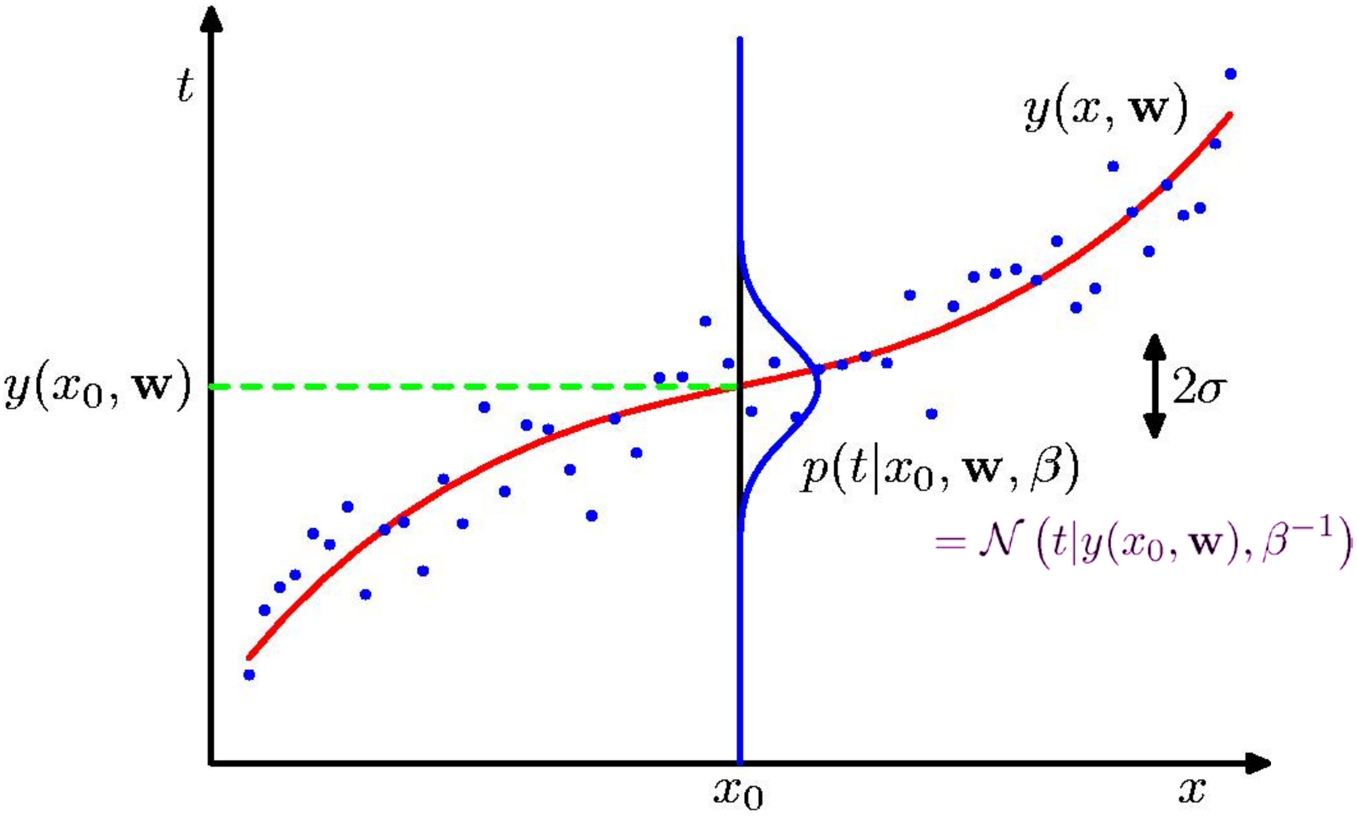
     (use same distribution over example space as training set)

Learning curve = % correct on test set as a function of training set size



Regression with Polynomials

Curve Fitting Re-visited



Maximum Likelihood

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

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

Determine ... ares error, $E(\mathbf{w})$

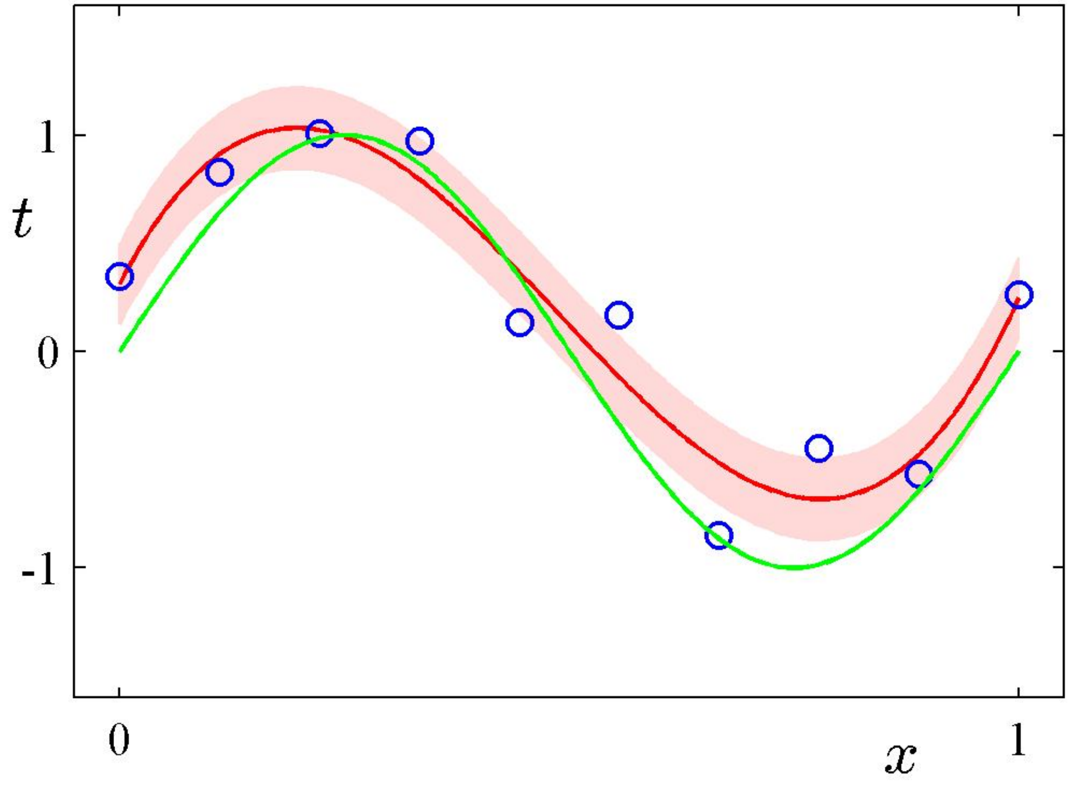
.

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

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Predictive Distribution

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



MAP: A Step towards Bayes

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

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

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

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$

.

MAP = Maximum Posterior

Bayesian Curve Fitting

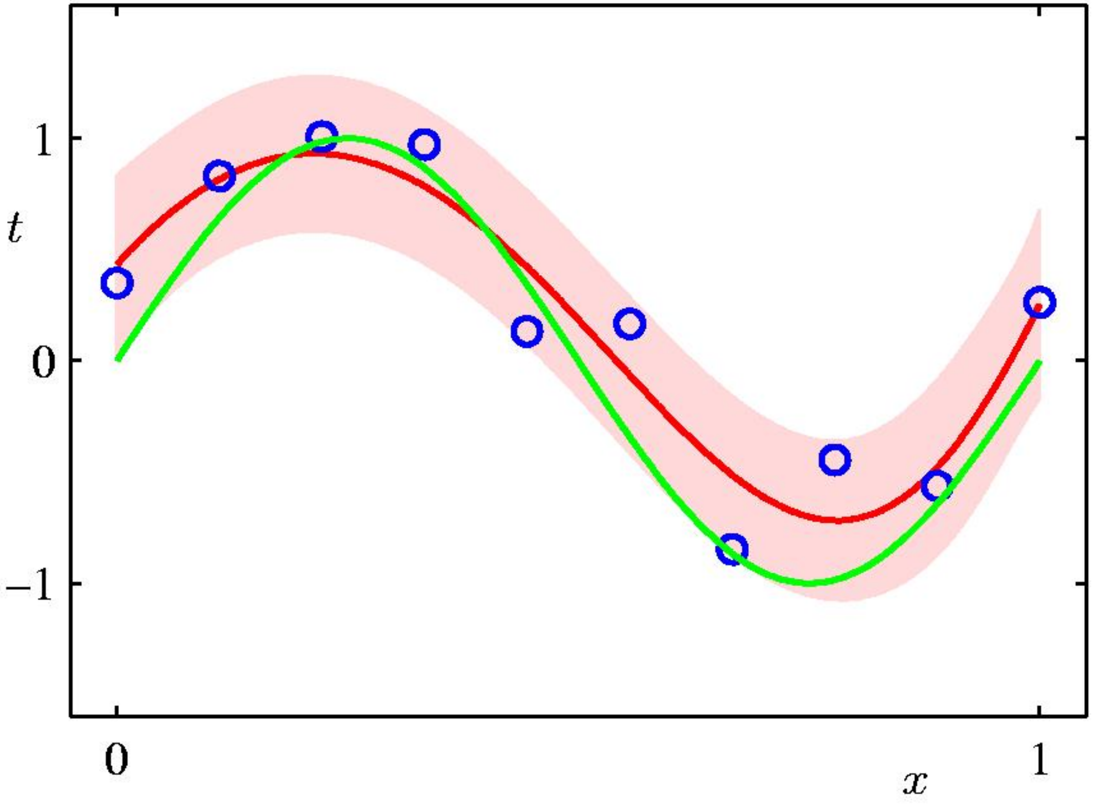
$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

Bayesian Predictive Distribution

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



Information Theory

Twenty Questions

Knower: thinks of object (point in a probability space)
Guesser: asks knower to evaluate random variables

Stupid approach:

Guesser: Is it my left big toe?
Knower: No.

Guesser: Is it Valmiki?
Knower: No.

Guesser: Is it Aunt Lakshmi?
...

Expectations & Surprisal

Turn the key: expectation: lock will open

Exam paper showing: could be 100, could be zero.
random variable: function from set of marks to real interval [0,1]

Interestingness $\propto$ unpredictability

$$\textit{surprisal}\ (r.v. = x) = -\log_2 p(x)$$

= 0 when p(x) = 1
= 1 when p(x) = ½
= ∞ when p(x) = 0

Expectations in data

A: 00010001000100010001. . . 0001000100010001000100010001

B: 01110100110100100110. . . 1010111010111011000101100010

C: 00011000001010100000. . . 0010001000010000001000110000

Structure in data → easy to remember

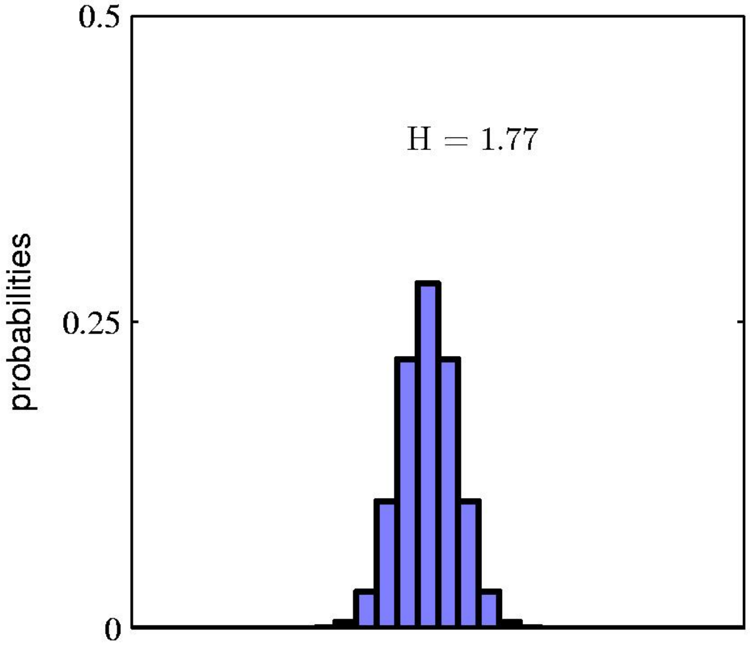
Entropy

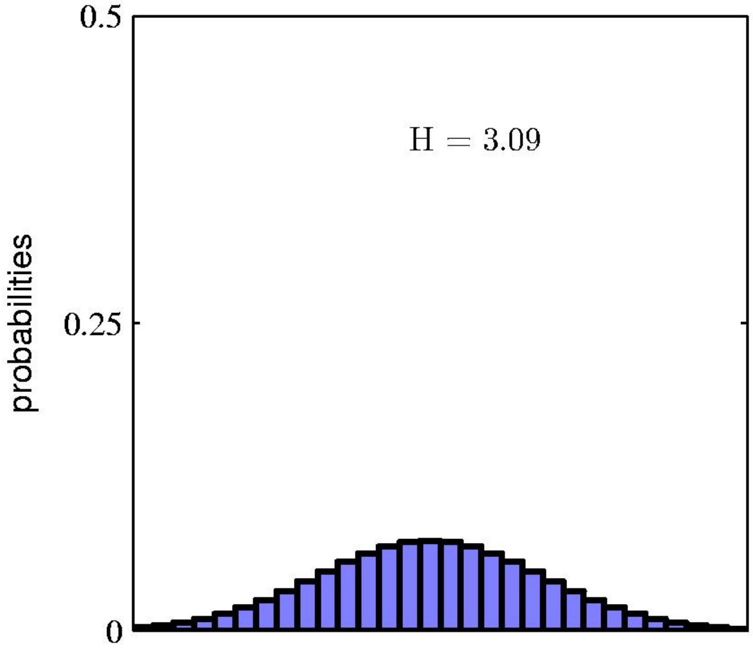
$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Used in

- coding theory
- statistical physics
- machine learning

Entropy





Entropy

In how many ways can N identical objects be allocated M bins?

$$W = \frac{N!}{\prod_i n_i!}$$

$$\mathrm{H} = \frac{1}{N} \ln W \simeq -\lim_{N \to \infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

Entropy maximized when $\forall i : p_i = \frac{1}{M}$

Entropy in Coding theory

x discrete with 8 possible states; how many bits to transmit the state of x?

All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

Coding theory

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Entropy in Twenty Questions

Intuitively : try to ask q whose answer is 50-50

Is the first letter between A and M?

question entropy = p(Y)logp(Y) + p(N)logP(N)

For both answers equiprobable:

$$\text{entropy} = - ½ * \log_2(½) - ½ * \log_2(½) = 1.0$$

For P(Y)=1/1028

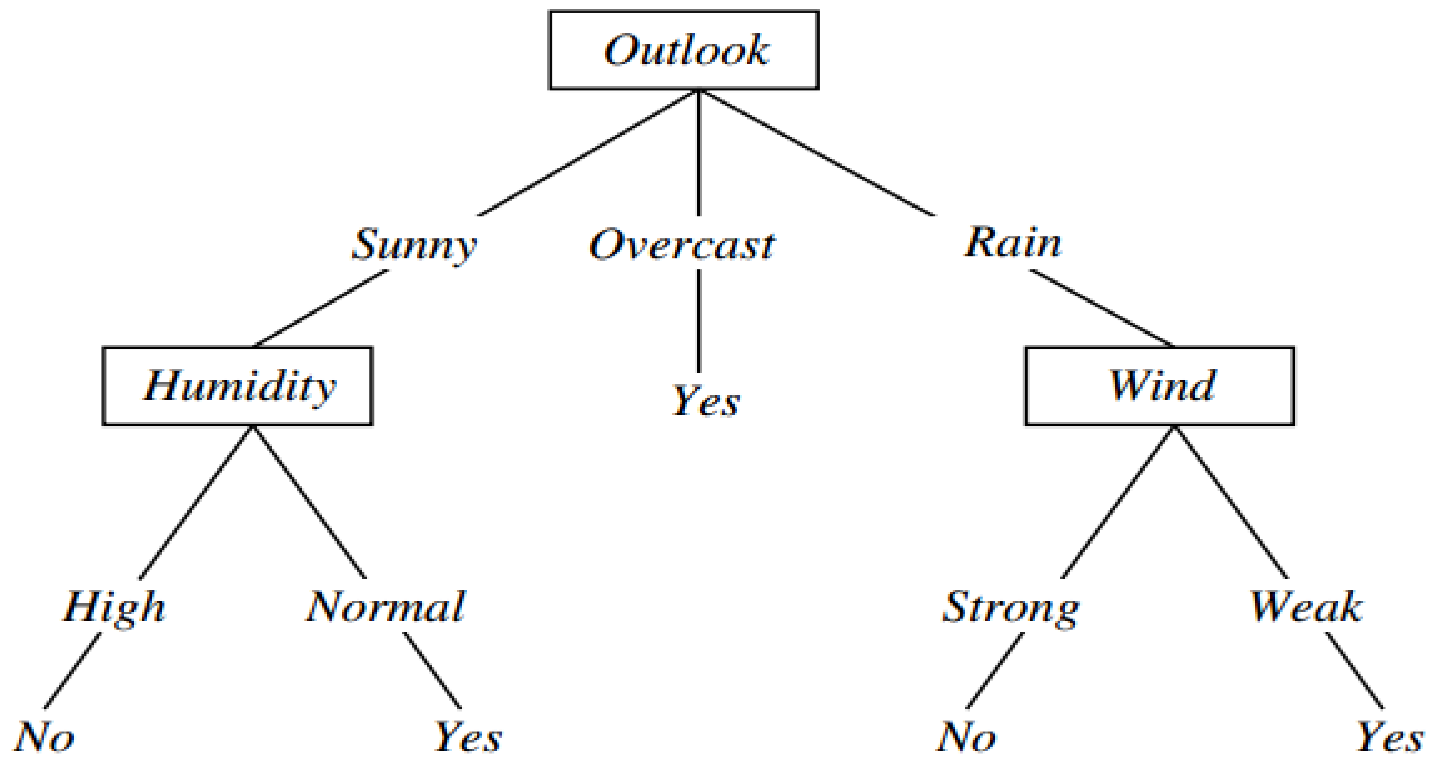
entropy = - 1/1028 * -10 - eps = 0.01

Learning Logical Rules
Decision Trees

Duda and Hart, Ch.1

Russell & Norvig Ch. 18

Boolean Decision Trees



Attribute-based representations

- Examples described by attribute values (Boolean, discrete, continuous)
- E.g., situations where I will/won't wait for a table:

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	\$\$\$	F	T	French	0–10	T
X_2	T	F	F	T	Full	\$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	\$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	\$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	\$\$\$	F	T	French	>60	F
X_6	F	T	F	T	Some	\$\$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	\$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	\$\$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	\$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	\$\$\$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	\$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	\$	F	F	Burger	30–60	T

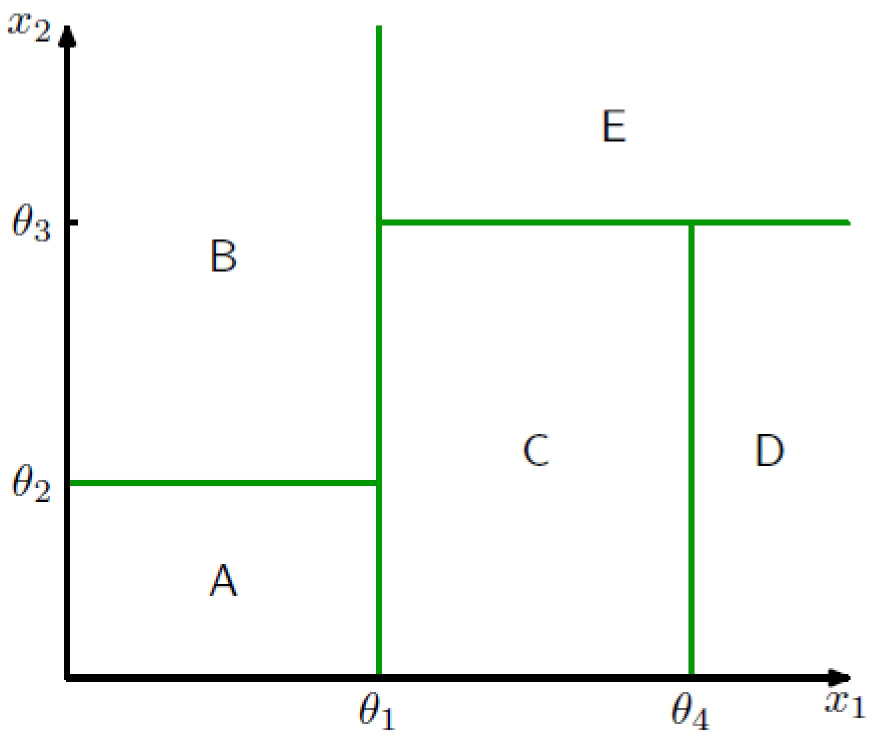
- Classification of examples is positive (T) or negative (F)

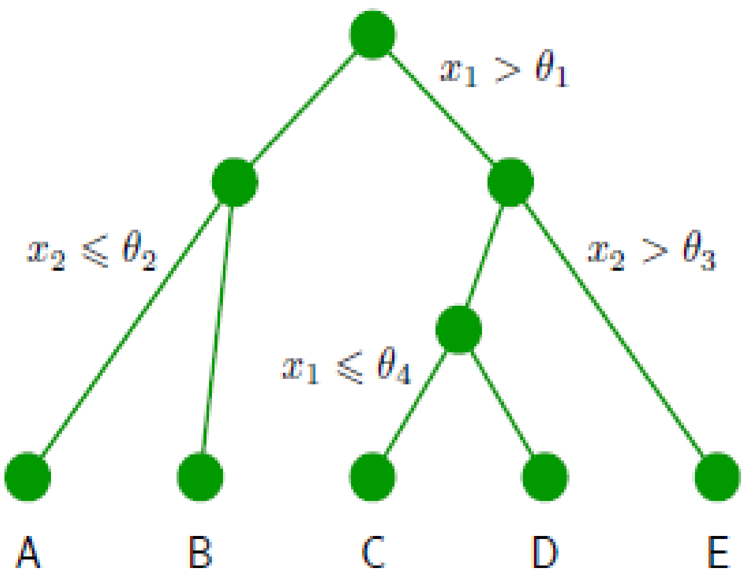
Learning decision trees

Problem: decide whether to wait for a table at a restaurant, based on the following attributes:

1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai, Burger)
10. WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

Continuous orthogonal domains





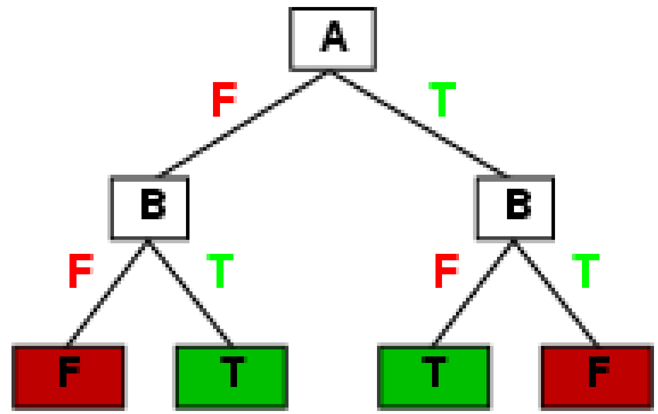
classification and regression trees CART

[Breiman 84]   ID3: [Quinlan 86]

Expressiveness

- Decision trees can express any function of the input attributes.
- E.g., for Boolean functions, truth table row → path to leaf:

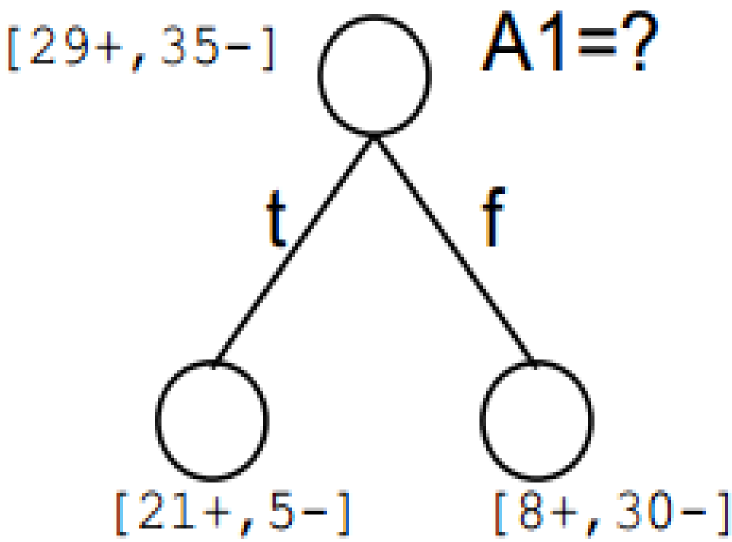
A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F

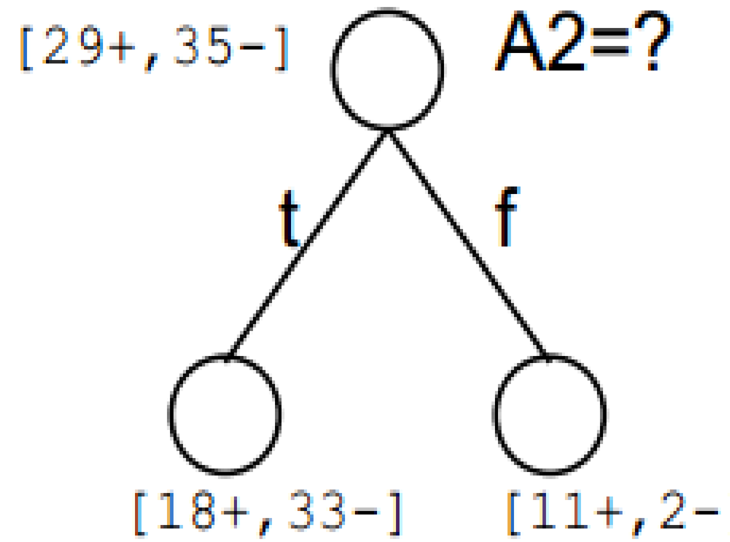


- Trivially, there is a consistent decision tree for any training set with one path to leaf for each example (unless *f* nondeterministic in *x*) but it probably won't generalize to new examples

- Prefer to find more compact decision trees

Which attribute to use first?

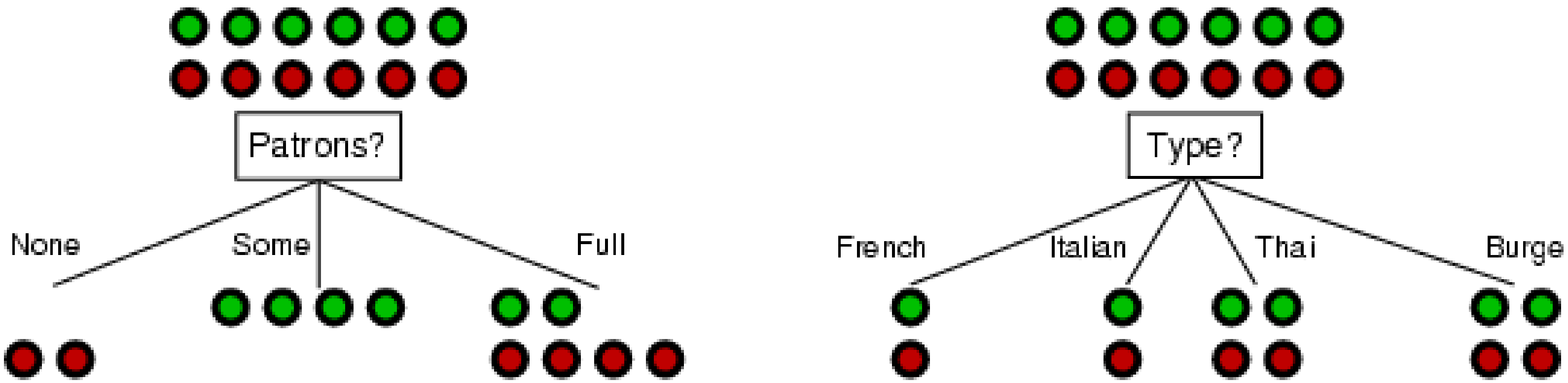




Gain(S; A) = expected reduction in entropy due to sorting on attribute A

Choosing an attribute

- **Idea: a good attribute splits the examples into subsets that are (ideally) "all positive" or "all negative"**



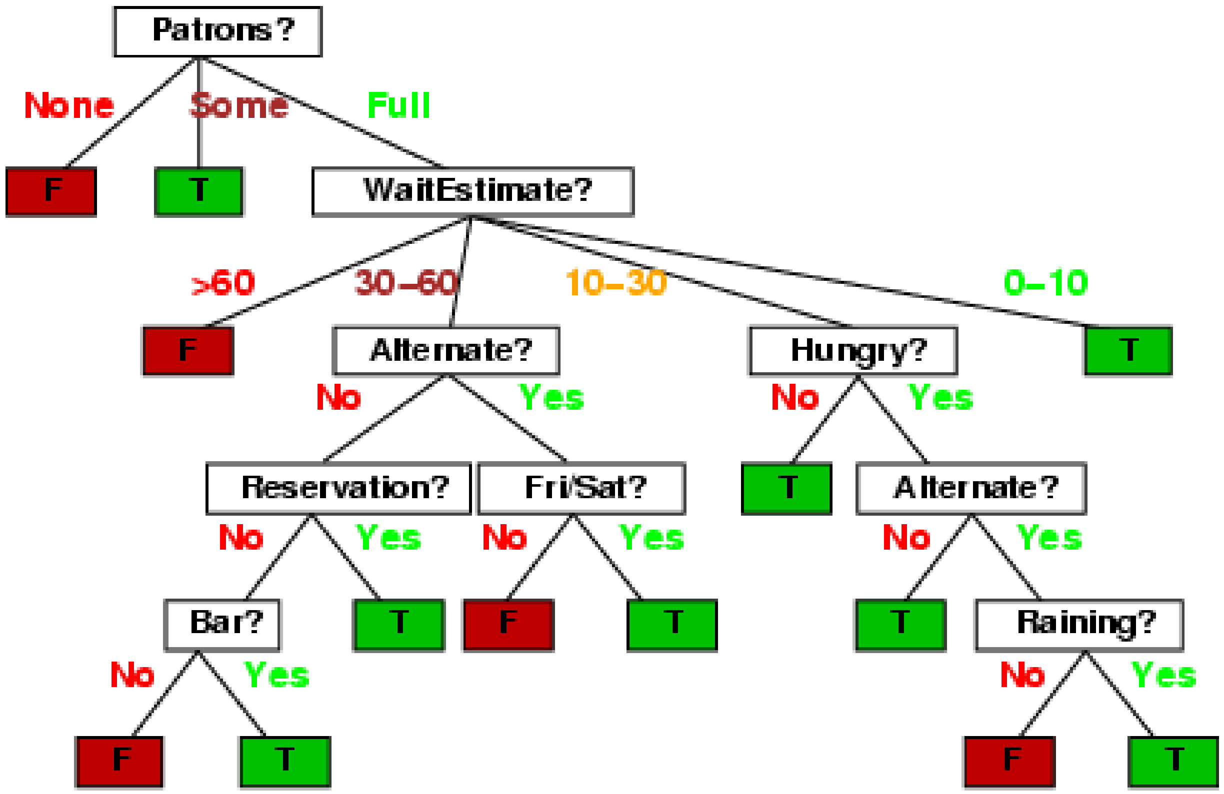
- ***Gain (Patrons)*** $= B(\frac{6}{12}) - [\frac{2}{12}B(\frac{0}{2}) + \frac{4}{12}B(\frac{4}{4}) + \frac{6}{12}B(\frac{2}{2})] = 0.541 bits$

- ***Gain (Type)*** $= 1 - [1.B(\frac{1}{2})] = 0$

Information Gain is higher for Patrons

Decision trees

- One possible representation for hypotheses
- E.g., here is the “true” tree for deciding whether to wait:



Information gain

- **A chosen attribute *A* divides the training set *E* into subsets $E_1, \dots, E_v$ according to their values for *A*, where *A* has v distinct values.**

$$remainder(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} B\left(\frac{p_i}{p_i + n_i}\right)$$

- **Information Gain (IG) or reduction in entropy from the attribute test:**

$$IG(A) = B\left(\frac{p}{p+n}\right) - remainder(A)$$

- **Choose the attribute with the largest IG**

Too many ways to order the tree

<u>How many distinct decision trees with *n* Boolean attributes?</u>

= number of Boolean functions

= number of distinct truth tables with 2^n rows = 2^{2^n}

- E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees

Hypothesis spaces

How many distinct decision trees with *n* Boolean attributes?

= number of Boolean functions

= number of distinct truth tables with 2^n rows = 2^{2^n}

- **E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees**

How many purely conjunctive hypotheses (e.g., *Hungry* $\wedge$ $\neg$*Rain*)?

- **Each attribute can be in (positive), in (negative), or out**

  $\Rightarrow$ 3^n distinct conjunctive hypotheses

- **More expressive hypothesis space**
  - increases chance that target function can be expressed
  - increases number of hypotheses consistent with training set

    $\Rightarrow$ may get worse predictions

Information gain

For the training set, $p = n = 6$, $I(6/12, 6/12) = 1$ bit

Consider the attributes *Patrons* and *Type* (and others too):

$$IG(Patrons) = 1 - [\frac{2}{12} I(0,1) + \frac{4}{12} I(1,0) + \frac{6}{12} I(\frac{2}{6}, \frac{4}{6})] = .541 \text{ bits}$$

$$IG(Type) = 1 - [\frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4})] = 0 \text{ bits}$$

Patrons **has the highest IG of all attributes and so is chosen by the DTL algorithm as the root**

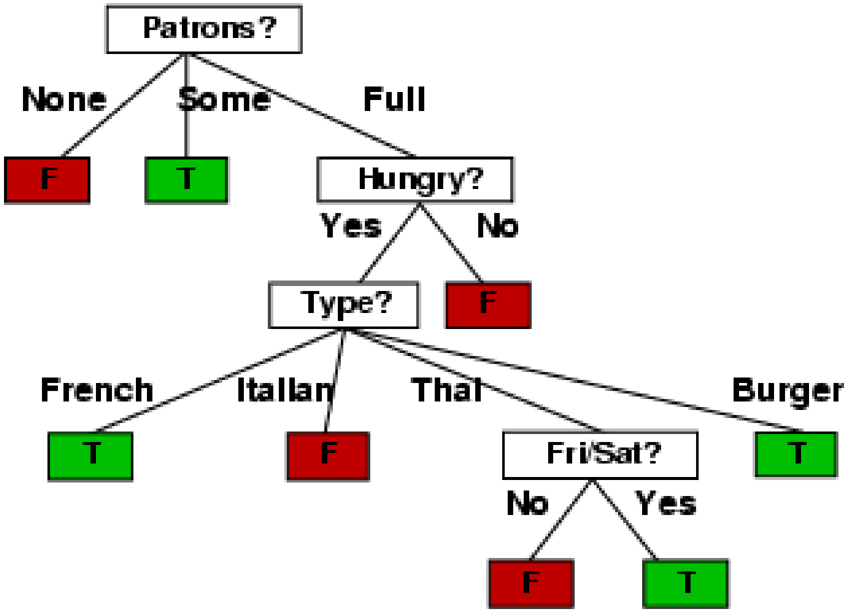
Decision tree learning

- **Aim: find a small tree consistent with the training examples**
- **Idea: (recursively) choose "most significant" attribute as root of (sub)tree**

```
function DTL(examples, attributes, default) returns a decision tree
    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value v_i of best do
            examples_i ← {elements of examples with best = v_i}
            subtree ← DTL(examples_i, attributes − best, MODE(examples))
            add a branch to tree with label v_i and subtree subtree
        return tree
```

Example contd.

- Decision tree learned from the 12 examples:



- Substantially simpler than “true” tree---a more complex hypothesis isn’t justified by small amount of data

Decision Theory

Decision Theory

Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

For given x, determine optimal t.

Minimum Expected Loss

Example: classify medical images as ‘cancer’ or ‘normal’

Loss matrix L:

		Decision: cancer	Decision: normal
Truth	cancer	0	1000
	normal	1	0

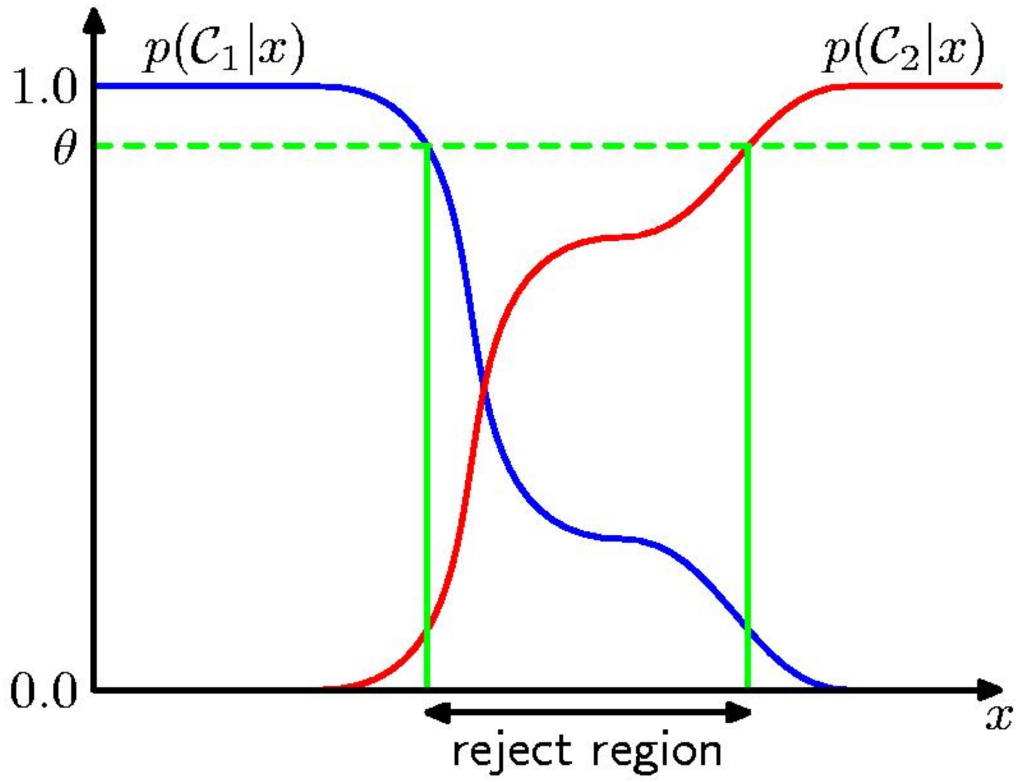
Minimum Expected Loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

Reject Option



Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Decision Theory for Regression

Inference step

Determine $p(\mathbf{x}, t)$.

Decision step

For given x, make optimal prediction, y(x), for t.

Loss function: $\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$

The Squared Loss Function

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

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

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \mathrm{var}\left[t|\mathbf{x}\right] p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$$