Introduction

Probability theory

A key concept in the field of pattern recognition is that of uncertainty. It arises both through noise on measurements, as well as through the finite size of data sets. Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition.

Decision theory

Combined with probability theory, it can allow us to make optimal decisions in situations involving uncertainty.

Information theory

A key measure in information theory is "entropy". Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process.

training set

target vector

training phase/learning phase

test set

preprocessed/ feature extraction

supervised learning

classification/regression

Unsupervised learning

clustering/density estimation

Reinforcement learning

Chapter 1 Introduction

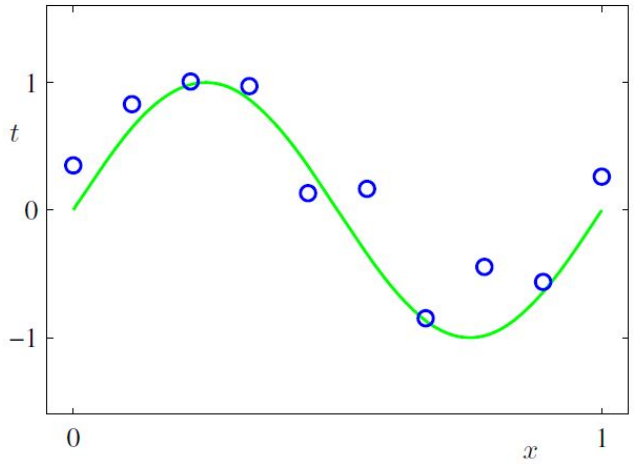
Example: Curve fitting problem

Training set:

$$\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$$ Input value

$$\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$$ Target value

Goal: Given a new $\hat{x}$, make the prediction of $\hat{t}$



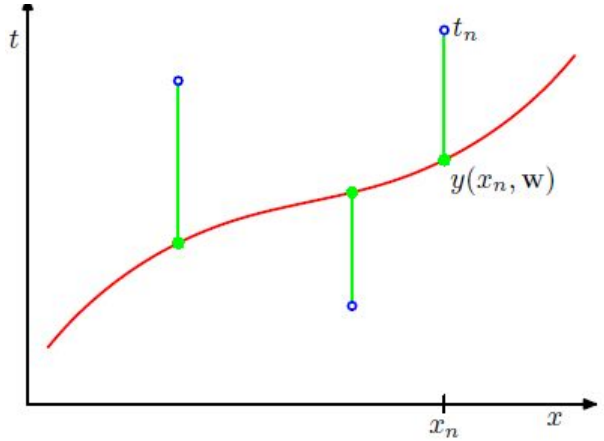
Chapter 1 Introduction

Solution 1: Error minimization

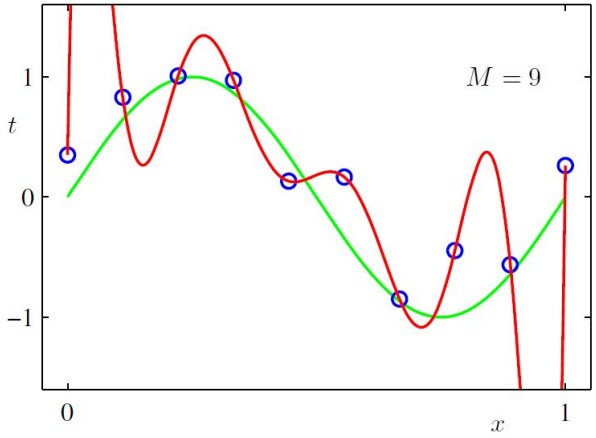
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

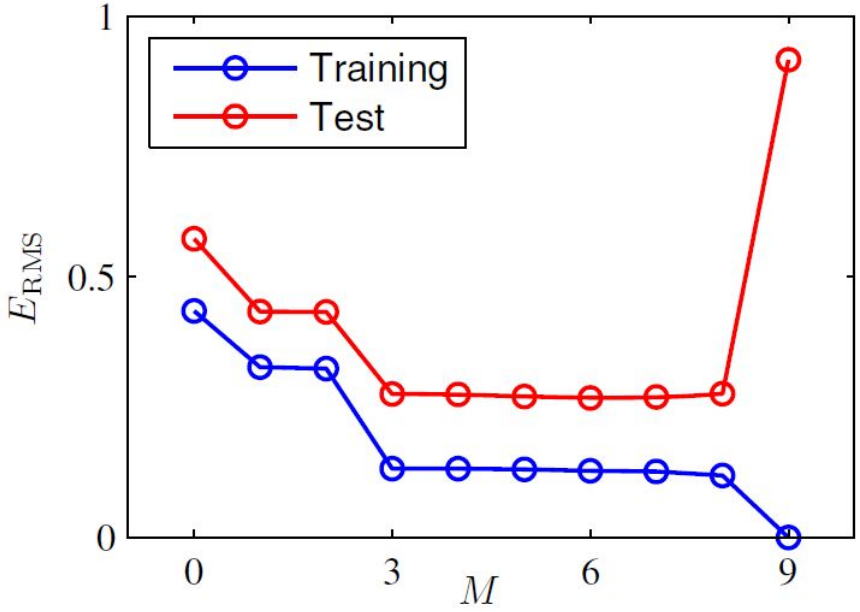
Minimize $$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

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



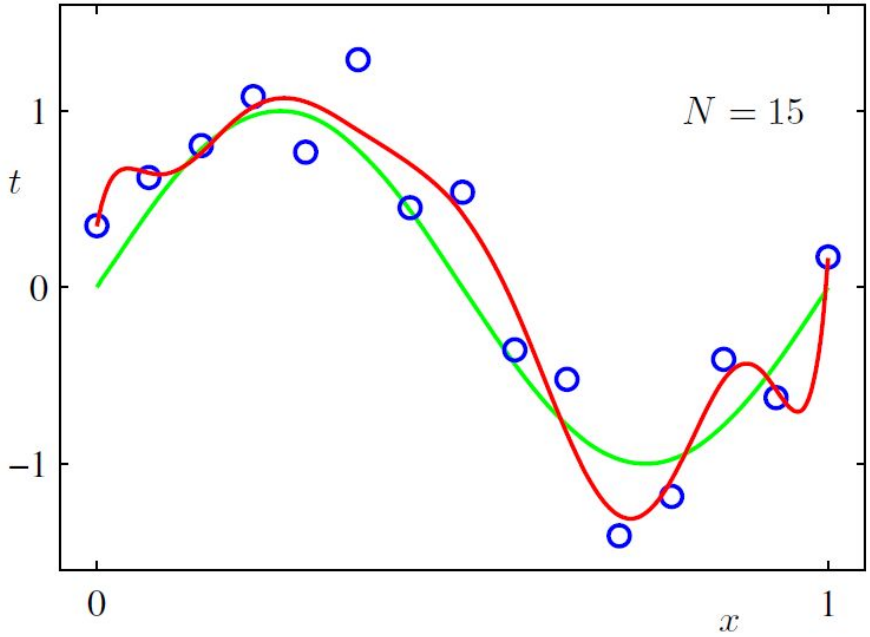
Over fitting

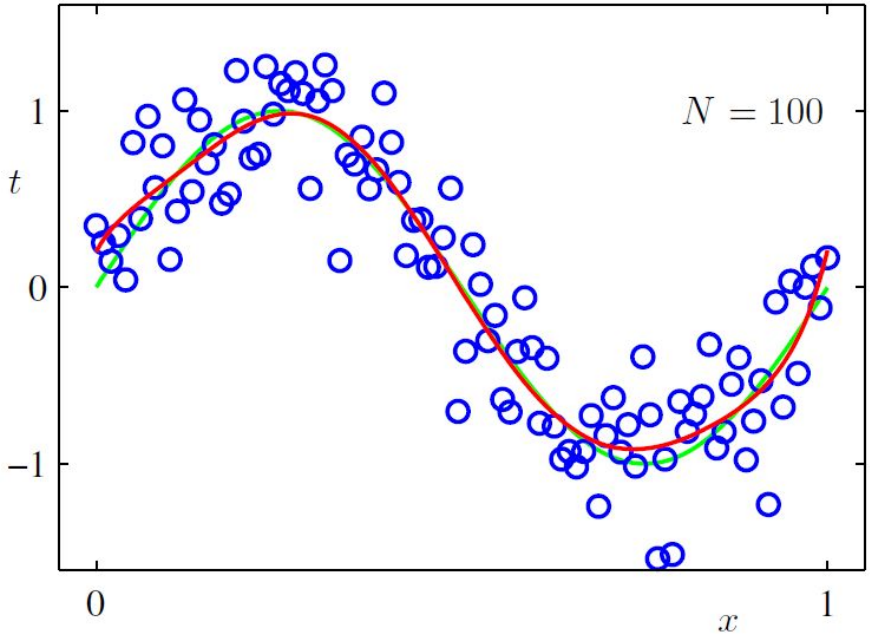




	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Increase sample size



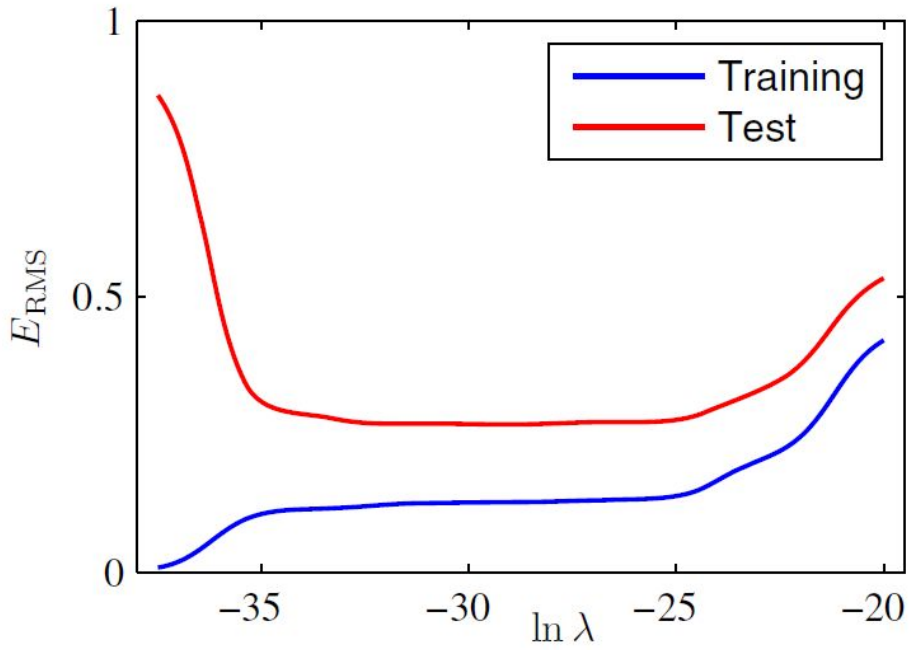


$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Shrinkage

NN:

Weight decay



Chapter 1 Introduction

1.1 Probability theory

Two fundamental rules of probability theory

sum rule $$p(X) = \sum_{Y} p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X).$$

Bayes' theorem

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

with $$p(X) = \sum_{Y} p(X|Y)p(Y).$$

Probability densities (continuous random variables)

$$p(x \in (a, b)) = \int_a^b p(x)dx$$

$$p(x) \geq 1,\ p(x \in (-\infty, \infty)) = 1$$

cumulative distribution function: $P(z) = p(x \in (-\infty, z))$

sum and product rules extend to probability densities.

Expectation: the average value of some function $f(x)$ under a probability distribution $p(x)$;

discrete case: $E[f] = \sum_x p(x)f(x)$

continuous case: $E[f] = \int p(x)f(x)dx$

N points *drawn from the prob. distribution or prob. density*, expectation can be approximated by:

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

Variance of $f(x)$: a measure of the variations of $f(x)$ around $E[f]$.

$$var[f] = E[f^2] - E[f]^2$$

$$var[x] = E[x^2] - E[x]^2$$

Covariance for two random variables:

$$cov[x, y] = E_{x,y}[xy] - E[x]E[y]$$

Two vectors of random variables:

$$cov[\boldsymbol{x}, \boldsymbol{y}] = E_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - E[\boldsymbol{x}]E[\boldsymbol{y}^\top]$$

Bayesian probabilities:
Frequentist estimator: maximum likelihood (MLE or ML);
Bayesian estimator: MLE and maximum a posteriori (MAP);

Bayesian:
the inclusion of prior knowledge arises naturally.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

This describes the uncertainty in model parameters.

$$p(\mathcal{D}) = \int \ldots \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Frequentist estimator: maximum likelihood (MLE or ML);

$$w^* = \max_w p(\mathcal{D}|w)$$

Uncertainty of w*?

Bootstrap

Bayesian: posterior distribution

Gaussian distribution

The Gaussian distribution of a single real-valued variable x:

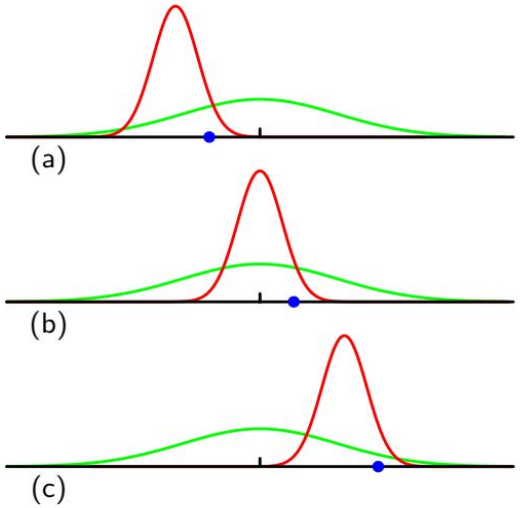
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

in D dimensions: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) : \mathbb{R}^D \rightarrow \mathbb{R}$
$E[x] = \mu$, var$[x] = \sigma^2$
$\mathbf{x} = (x_1,\ldots,x_N)$ is a set of N observations of the SAME scalar variable x
Assume that this data set is *independent and identically distributed*:

$$p(x_1,\ldots,x_N|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

maximum likelihood solution :

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



$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n$$

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

Bias

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

Curve fitting re-visited

MLE

$\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$

$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

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

$$\beta^{-1} = \sigma^2$$

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

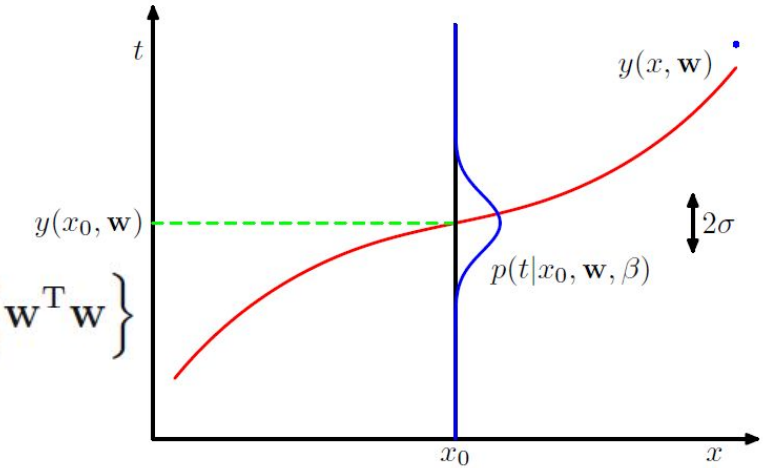
$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

MAP

Assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x,\ w)$.

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

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



A more Bayesian approach: introduce a prior distribution over the polynomial coefficients **w**.

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

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

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

Equivalent to the formal solution

Assume that the parameters α and β are fixed and known in advance.

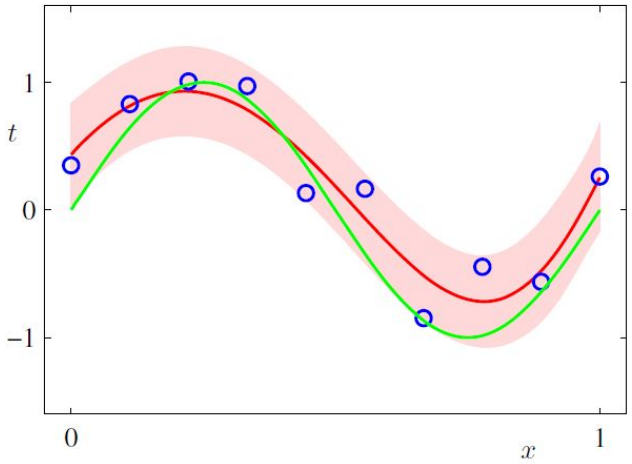
In a fully Bayesian approach, we should consistently apply the sum and product rules of probability, which requires that we integrate over all values of **w**.

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

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

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

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x).$$



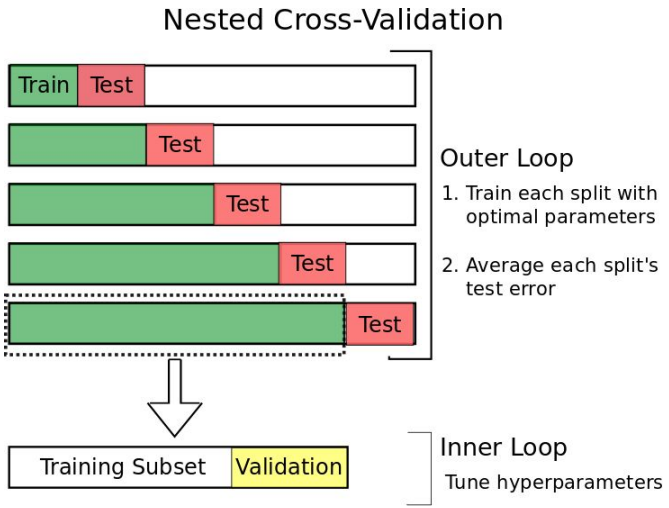
Model selection

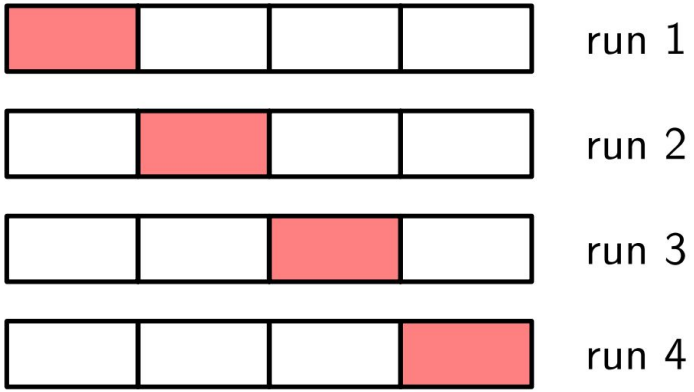
Training set

Validation set

Test set

cross-validation





Model selection

which is the optimal order of the polynomial that gives the best generalization?

train a range of models and test them on an independent *validation set*

cross-validation: use a subset for training and the whole set for assessing the performance

Akaike information criterion: $\ln p(\mathcal{D}|\boldsymbol{w}_{ML}) - M$

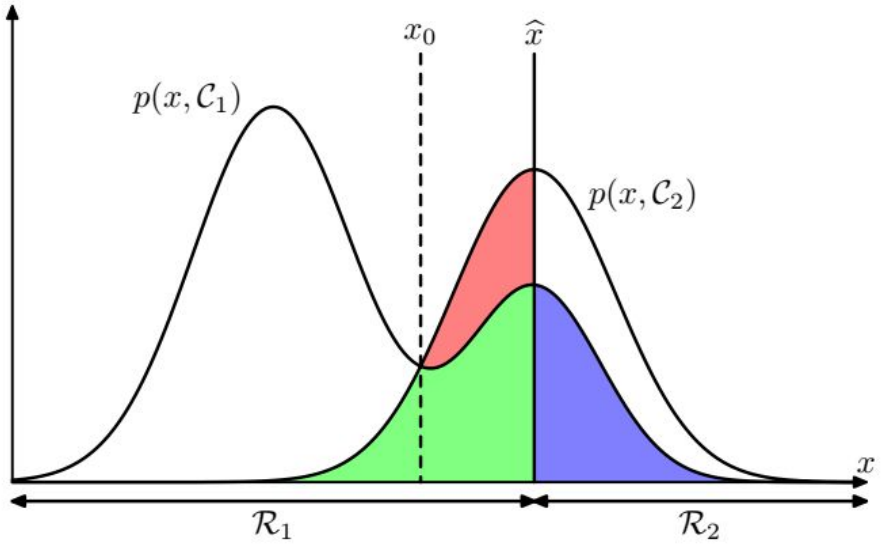
Bayesian information criterion (BIC)

The curse of dimensionality

Increasing number of features data density decreases and complexity increases and it became very difficult for machine learning model to work efficiently.

Minimizing the misclassification rate

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



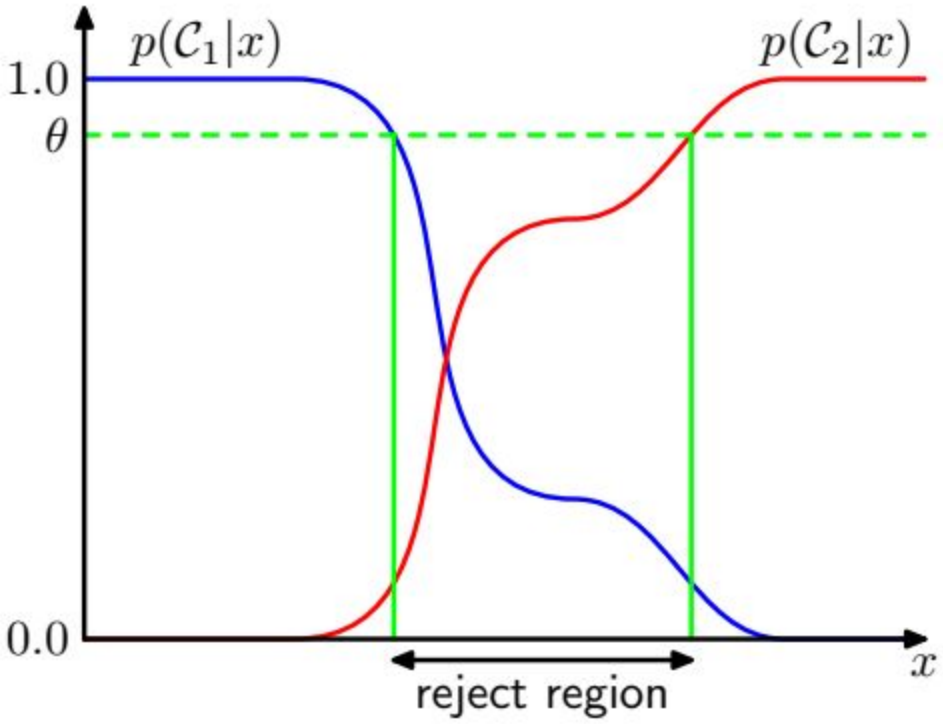
Minimizing the expected loss

$$\begin{array}{c} \\ \text{cancer} \\ \text{normal} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

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

minimize $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$

Reject region



inference stage -> train $p(\mathcal{C}_k|\mathbf{x})$

decision stage -> make optimal class assignments.

A. Generative models

1. determine the class-conditional densities
2. infer the prior class probabilities
3. Use Bayes' theorem to find the posterior class probabilities

(outlier detection)

B. Discriminative models

1. determine the posterior class probabilities
2. use decision theory to assign each new x to one of the classes.

C. Find a function f(x) In this case, probabilities play no role. A single learning problem.

Minimizing risk

Reject option

Compensating for class priors

Combining models

Regression example

Information theory introduction

- Consider a discrete random variable X
- We want to define a measure $h(x)$ of surprise/information of observing $X = x$
- Natural requirements:
  - if $p(x)$ is low (resp. high), $h(x)$ should be high (resp.low)
    - $h(x)$ should be a monotonically decreasing function of $p(x)$
  - if X and Y are unrelated, $h(x, y)$ should be $h(x) + h(y)$
    - i.e., if X and Y are independent, that is $p(x, y) = p(x)p(y)$

  $\Rightarrow$ this leads to $h(x) = -\log p(x)$
- Entropy of the variable X:

$$H[X] = E[h(X)] = -\sum_x p(x) \log(p(x))$$

$p(x) = 0?$

Consider a random variable x having 8 possible

states, each of which is equally likely. In order to communicate the value of x to

a receiver, we would need to transmit a message of length 3 bits.

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

a variable having 8 possible states {a, b, c, d, e, f, g, h}

(½,¼,⅛,1/16,1/64,1/64,1/64,1/64)

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

0, 10, 110, 1110, 111100, 111101, 111110, 111111

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

Stirling's approximation

$$\ln N! \simeq N \ln N - N$$

Some remarks:

- $H[X] \geq 0$ since $p \in [0, 1]$ (hence $p \log p \leq 0$)
- $H[X] = 0$ if $\exists x$ s.t. $p(x) = 1$
- Maximum entropy distribution $=$ uniform distribution
  - optimization problem: maximize $H[X] + \lambda\left(\sum_{x_i} p(x_i) - 1\right)$
  - derivating w.r.t. $p(x_i)$ shows they must be constant
  - hence $p(x_i) = 1/M, \forall x_i \Rightarrow H[X] = \log(M)$

  $\Rightarrow$ we therefore have $0 \leq H[X] \leq \log(M)$

- $H[X] =$ lower bound on the $\#$ of bits required to (binarily) encode the values of X (using $\log_2$ in the defintion of H)
  - trivial code of length $\log_2(M)$ (ex: $M = 8$, messages of size 3)
    - no "clever" coding scheme for uniform distributions
  - for non-uniform distributions, optimal coding schemes can be designed
    - high probability values $\Rightarrow$ short codes

Entropy, the average amount of information needed to specify the state of a random variable

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

Relative entropy

Unknown distribution $p(\mathbf{x})$

Approximating distribution $q(\mathbf{x})$

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

Mutual information

A measure of the mutual dependence between the two variables.

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \end{aligned}$$

KL

KL Divergence helps us to measure just how much information we lose when we choose an approximation.

$$LR = \frac{p(x)}{q(x)}$$

$$X = \{x_1, \ldots, x_N\}, \quad x_i \overset{i.i.d.}{\sim} p(x)$$

$$\widehat{\log LR}(X) = \frac{1}{N} \log \frac{p(x_1, \ldots, x_N)}{q(x_1, \ldots, x_N)}$$

$$= \frac{1}{N} \sum_{i=0}^{N} \log \frac{p(x_i)}{q(x_i)}$$

$$x_i \sim p(x)$$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N} \log \frac{p(x_i)}{q(x_i)} = E_{x \sim p(x)} \{ \log \frac{p(x)}{q(x)} \}$$

Data generated by an (unknown) distribution $p(x)$

We want to fit a parametric probabilistic model $q(x|\theta) = q_\theta(x)$

$\Rightarrow$ i.e., we want to minimize $KL(p||q_\theta)$

Data available: observations $(x_1, \ldots, x_N)$:

$$
\begin{aligned}
KL(p||q_\theta) &= -\ln \int p(x) \times \ln \frac{q(x|\theta)}{p(x)} dx \\
&\simeq -\sum_{i=1}^{N} \ln \frac{q(x_i|\theta)}{p(x_i)} \\
&= \sum_{i=1}^{N} \Big(-\ln q(x_i|\theta) + \ln p(x_i) \Big)
\end{aligned}
$$

$\Rightarrow$ it follows that minimizing $KL(p||q_\theta)$ corresponds to maximizing $\sum_{i=1}^{N} \ln q(x_i|\theta) =$ log-likelihood

Kullback-Leibler divergence between distributions p and q:

$$KL(p||q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx \right)$$

$$= -\int p(x) \ln \frac{q(x)}{p(x)} dx$$

$KL(p||q) \neq KL(q||p)$

$KL(p||p) = 0$

$KL(p||q) \geq 0$

Information theory - Mutual information

- Mutual information: $I[X,Y] = KL(p(X,Y)||p(X)_0 p(Y))$
- Quantifies the amount of independence between X and Y
  - $I[X,Y] = 0 \Leftrightarrow p(X,Y) = p(X)p(Y)$
- We have:

$$
\begin{aligned}
I[x,y] &= -\int\int p(x,y) \ln \frac{p(x)p(y)}{p(x,y)} dxdy \\
&= -\int\int p(x,y) \ln \frac{p(x)p(y)}{p(x|y)p(y)} dxdy \\
&= -\int\int p(x,y) \ln \frac{p(x)}{p(x|y)} dxdy \\
&= -\int\int p(x,y) \ln p(x) dxdy - \Big(-\int\int p(x,y) \ln p(x|y) dxdy \Big) \\
&= H[X] - H[X|Y]
\end{aligned}
$$

- Conclusion: $I[X,Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$
  - $I[X,Y]$ = reduction of the uncertainty about X obtained by telling the value of Y (that is, 0 for independent variables)

($\star\star$) **WWW** Consider two variables $\mathbf{x}$ and $\mathbf{y}$ having joint distribution $p(\mathbf{x}, \mathbf{y})$. Show that the differential entropy of this pair of variables satisfies

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] \leqslant \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] \tag{1.152}$$

with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are statistically independent.