

Chapter 1 : Introduction

sugyeong jo

2019 8 26

Contents

1.0 Introduction

1.1 Example: Polynomial Curve Fitting

1.2 Probabililty Theory

- 1.2.1 Probability densities
- 1.2.2 Expectations and covariabces
- 1.2.3 Bayesian probabilities
- 1.2.4 The Gaussian distribution
- 1.2.5 Curve fitting re-visited
- 1.2.6 Bayesian curve fitting

1.3 Model Selection

1.4 The Curse of Dimensionality

1.5 Decision Theory

- 1.5.1 Minimizaing the misclassification rate
- 1.5.2 Minimizing the expected loss
- 1.5.3 The reject option
- 1.5.4 Inference and decision
- 1.5.5 Loss functions for regression

1.6 Information Theory

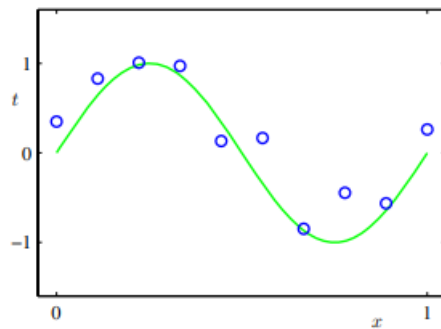
- 1.6.1 Relative entropy and mutual information

1. Introduction

- Training set: $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, to tune the parameters of an adaptive model
- Target vector: \mathbf{t} , the identity of the corresponding training set digit.
- Training phase / Learning phase / Generalization
- Preprocessing: to transform the original input variables into some new space of variables \rightarrow easy to recognition pattern
 - purpose: (dimensionality reduction \rightarrow) (1) feature extraction, (2) to speed up computation
- Application
 - supervised learning: classification, regression
 - unsupervised learning: clustering, density estimation
 - reinforcement learning: finding suitable actions to take in a given situation in order to maximize a reward.

1.1 Example: Polynomial Curve Fitting

Goal: to exploit the training set in order to make predictions of the value \hat{t} of the target variable for some new value \hat{x} of the input variable.



- trying to discover the underlying function $\sin(2\pi x)$

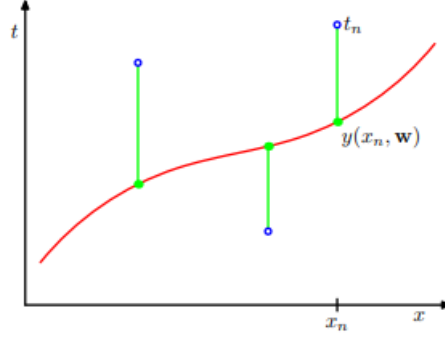
$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1.1)$$

- M is order of the polynomial
- $y(x, \mathbf{w})$ is a nonlinear function of x and a linear function of the coefficients \mathbf{w} .

Step 1: choosing the value of \mathbf{w} to minimize error function, $E(\mathbf{w})$.

- error function

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad (1.2)$$



- by the sum of the squareds of the errors between the predictions $y(x_n, \mathbf{w})$ for each data point x_n and the corresponding target value t_n
- 1/2: for convenience
- result is positive quantity and that would be zero, iff the function $y(x_n, \mathbf{w})$ were to pass exactly through each training data point.

(ex.1.1)

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\} x_n^i = 0$$

$$\sum_{n=1}^N y(x_n, \mathbf{w}) x_n^i = \sum_{n=1}^N t_n x_n^i$$

$$\sum_{n=1}^N \left(\sum_{j=0}^M w_j x_n^j \right) x_n^i = \sum_{n=1}^N t_n x_n^i$$

$$\sum_{n=1}^N \sum_{j=0}^M w_j x_n^{(j+i)} = \sum_{n=1}^N t_n x_n^i$$

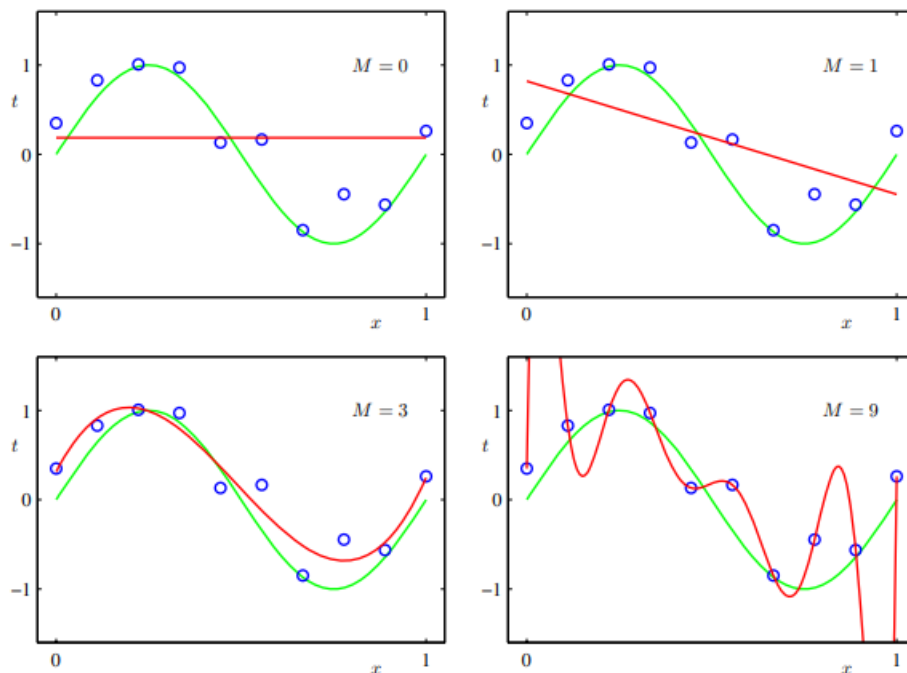
$$\sum_{j=1}^M \sum_{n=0}^N x_n^{(j+i)} w_j = \sum_{n=1}^N t_n x_n^i$$

$$\sum_{j=1}^M A_{ij} w_j = T_i$$

\therefore the jcoefficients \mathbf{w} that minimize the error function are given by the solution to above set of linear equations.

Step 2: choosing the order M of polynomial (*model comparison* or *model selection*)

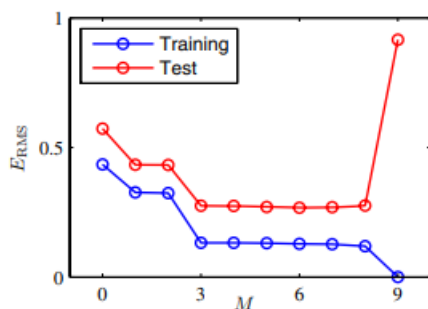
- over-fitting: eventhough, when polynomial passes exactly through each data point ($M = 9$), error function is 0, $E(\mathbf{w}^*) = 0$, the fitted curve oscillates widly and gives a very poor representation of the function $\sin(2\pi x)$.



- root mean square error, RMS error

$$E_{RMS} = \sqrt{2E(\mathbf{w}^*)/N} \quad (1.3)$$

- the division by N : it could be to compare different sizes of data sets on an equal footing
- square root: measured on the same scale as the target variable t .

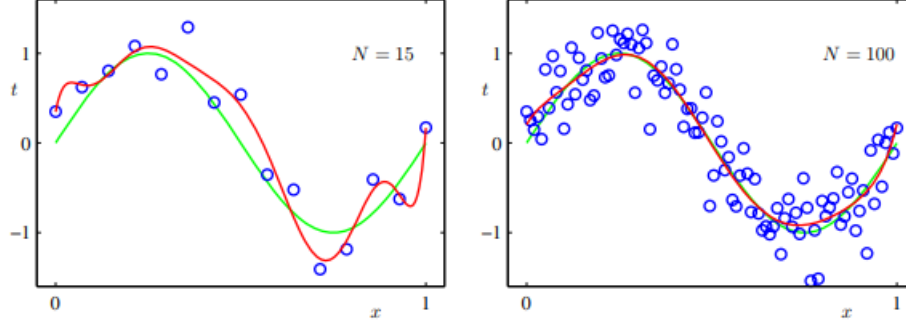


	$M = 0$	$M = 1$	$M = 6$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

- (*over-fitting problem*) larger values of $M \rightarrow$ the more flexible, increasing the coefficients $\mathbf{w}^* \rightarrow$ increasingly tuned to the random noise on the target values

- over come the over-fitting problem

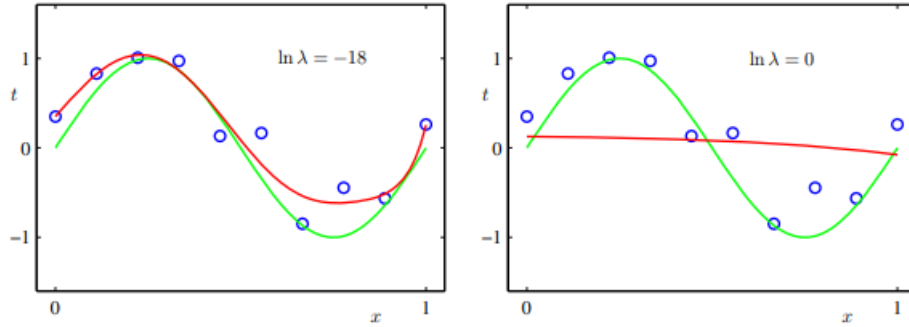
- increasing the size of data set
- Bayesian method
- (for limit size) Regularization: adding the penalty term to the error function (1.2)



- Regularization

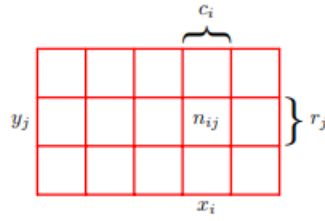
$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad (1.4)$$

- λ : reactive importance of the regularization term compared with the sum-of-squares error term.
 - zero: overfitting \rightarrow desired: good for fitting \rightarrow too large: poor fit
- w_0 : normally omitted from the regularizer (\because it depend on the choice of origin for the target variable)
- shrinkage method (e.g. ridge regression, weight decay, ...)



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

1.2 Probabililty Theory



- joint probability:

$$p(X = x_i, Y = y_i) = \frac{n_{ij}}{N} \quad (1.5)$$

- sum rule (marginal probability):

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_{j=1}^L p(X = x_i, Y = y_j) \quad (1.7)$$

- condition probability:

$$p(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i} \quad (1.8)$$

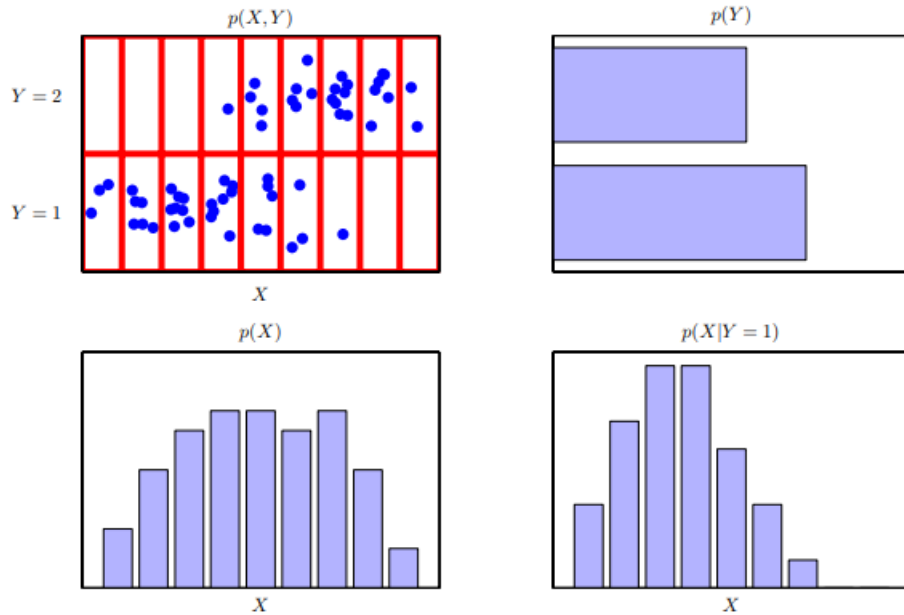
- product rule:

$$p(X = x_i, Y = y_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_{ij}} \cdot \frac{c_{ij}}{N} = p(Y = y_i | X = x_i) p(X = x_{ij}) \quad (1.9)$$

- Bayes' theorem

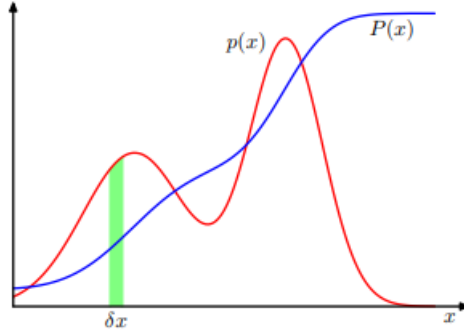
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)} \quad (1.13)$$

- independent: $p(X, Y) = p(X)p(Y)$



1.2.1 Probability densities

- **probability density:** If the probability of a real-valued variable x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \rightarrow 0$, then $p(x)$ is probability density.



$$p(x \in (a, b)) = \int_a^b p(x) dx \quad (1.24)$$

s.t

$$p(x) \geq 0 \quad (1.25)$$

$$\int_{-\infty}^{\infty} p(x) dx = 1 \quad (1.26)$$

- probability density transforms (due to Jacobian factor)
 - Jacobian factor: $J_{ki} \equiv \frac{\partial y_k}{\partial x_i}$
 - if $x = g(y)$, $f(x) = \tilde{f}(y) = f(g(y)) \rightarrow p_y(y) \neq p_x(x)$
 - (ex.1.4) the concept of the maximum of a probability density is dependent on the choice of variable.

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| = \frac{d(p_x(g(y))|g'(y)|)}{dy} \quad (1.27)$$

maximum value is calculated by $dp_y(y)/dy|_{\hat{y}} = 0$

$$\frac{dp_y(y)}{dy} = \frac{d(p_x(g(y))|g'(y)|)}{dy} \quad (1)$$

$$= \frac{d(p_x(g(y)))}{dy} |g'(y)| + p_x(g(y)) \frac{d|g'(y)|}{dy} \quad (2)$$

$$= \frac{dp_x(g(y))}{dg(y)} \frac{dg(y)}{dy} |g'(y)| + p_x(g(y)) \frac{d|g'(y)|}{dy} \quad (3)$$

$$= \frac{dp_x(x)}{dx} \frac{dg(y)}{dy} |g'(y)| + p_x(g(y)) \frac{d|g'(y)|}{dy} = p_x(g(y)) \frac{d|g'(y)|}{dy} \quad (4)$$

If $p_x(x) = 2x$, $x \in [0, 1]$, the maximum value of variable \hat{x} is 1. And given that $x = \sin(y)$, it transform to the $p_y(y) = 2\sin(y)|\cos(y)| (= \sin(2y))$, $y \in [0, \pi/2]$, and the \hat{y} is $\pi/4$. $\therefore \hat{x} \neq \sin(\hat{y})$

- **cumulative distribution function:** the probability that x lies in the interval $(-\infty, z)$
- **probability mass function :** $p(x)$ when x is a discrete variable.

$$P(z) = \int_{-\infty}^z p(x)dx \quad (1.28)$$

- The sum and product rules, Bayes' theorem of probability densities

$$p(x) = \int p(x, y)dy \quad (1.31)$$

$$p(x, y) = p(y|x)p(x) \quad (1.32)$$

1.2.2 Expectations and covariabces

- **expectation of $f(x)$:** weighted by the relative probabilities of the different values of x .

$$\mathbb{E}[f] = \sum_x p(x)f(x) \text{ or } \int p(x)f(x)dx \quad (1.33)$$

- if there is N of points, the expectation can be approximated as

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n) \quad (1.35)$$

- **conditional expectation**

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \quad (1.37)$$

- **variance:** measurment of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.

$$\begin{aligned} var[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)\mathbb{E}[f(x)]] + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \end{aligned} \quad (1.38)$$

$$var[f] = \mathbb{E}[(f(x))^2] - \mathbb{E}[f(x)]^2 \quad (1.39)$$

- **covariance:** expresses the extent to which x and y vary together. (if x and y is independent, cov=0)

$$\begin{aligned} cov[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy - x\mathbb{E}[y] - y\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y]] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}_{x,y}[x\mathbb{E}[y]] - \mathbb{E}_{x,y}[y\mathbb{E}[x]] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \quad (1.41)$$

(vector)

$$\begin{aligned} cov[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned} \quad (1.42)$$

1.2.3 Bayesian probabilities

- **Purpose:** to address and quantify the uncertainty that surrounds the appropriate choice for the model parameters \mathbf{w}
- **Bayes' theorem:** at the uncertain event,
 - (1) (prior probability) Suppose some opinion based on exist knowldge
 - (2) obtain fresh evidence
 - (3) (posterior probability) revise the uncertainty about (1)opinion
 - that is, to convert a prior probability into a posterior probability by incorporating the evidence provided by the observed data

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})} \quad (1.43)$$

- $p(\mathbf{w})$: prior probability, assumptions about \mathbf{w} , before observing the data
- $p(\mathcal{D}|\mathbf{w}), \mathcal{D} = \{t_1, t_2, \dots, t_n\}$: likelihood function, how probable the observed data set is for different settings of the parameter vector \mathbf{w}
- $p(\mathbf{w}|\mathcal{D})$: posterior probability, to evaluate the uncertainty in \mathbf{w} after observing \mathcal{D} .
- $p(\mathcal{D})$: normalization constant
- **Frequentist paradigms**
 - frequentist estimator \rightarrow maximum likelihood (maximize $p(\mathbf{w}|\mathcal{D})$)
 - or minimize the *error* by the *error function*
- **Bayesian view:** provide a quantification of uncertainty using probabilities.
 - Advantage: the inclusion of prior knowledge arises naturally
 - Criticism: at the prior distribution is often selected on the basis of mathematical convenience rather than as a reflection of any prior beliefs
 - To reduce the dependence on the prior \rightarrow noninformative priors
 - Limitation: for using Bayeesian, need to marginalize over the whole of parameter space (it is difficult!)

1.2.4 The Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (1.46)$$

- precision: $\frac{1}{\sigma^2}$

- **vector form**

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (1.52)$$

It is probability density

$$(1) \mathcal{N}(x|\mu, \sigma^2) > 0$$

$$(2)$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \quad (1.48)$$

(ex.1.7)

$$\begin{aligned} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} dx &= \sqrt{2\pi\sigma^2} \\ \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{x - \mu}{\sqrt{2\sigma^2}}\right)^2\right\} dx &= \sqrt{2\pi\sigma^2} \\ \int_{-\infty}^{\infty} e^{-z^2} dz &= \sqrt{\pi}, \text{ where } z = \left(\frac{x - \mu}{\sqrt{2\sigma^2}}\right), dx = \sqrt{2\sigma^2} dz \end{aligned}$$

transform the spherical coordinate system

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy &= \int \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy = \pi \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta = \int_0^{2\pi} \int_0^{\infty} (-1/2) e^u du d\theta = \pi \end{aligned}$$

- **expectation**

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu \quad (1.49)$$

(ex.1.8)

$$\begin{aligned} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx &= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x dx \\ &= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} (y+\mu) dy \quad (y = x - \mu) \\ &= \mu \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} dy + \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} y dy \\ &= \mu + 0 = \mu \end{aligned}$$

- **variance**

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2 \quad (1.50)$$

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \quad (1.51)$$

- The maximum of a distribution is known as its mode. For a Gaussian, the mode = mean

(ex.1.8)

$$\begin{aligned} \text{Var}[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu, \sigma^2) x dx, f(x) = \mathcal{N}(x|\mu, \sigma^2) \\ &= \int_{-\infty}^{\infty} x^2 f(x) - 2\mu x f(x) + \mu^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \cdot \mu + \mu^2 \cdot 1 \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \end{aligned}$$

$$\therefore \mathbb{E}[x^2] = \text{Var}[x] + \mu^2 = \sigma^2 + \mu^2$$

Goal: determine μ, σ parameters from the data set

- **maximize the (log) likelihood function**

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{1}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \quad (1.54)$$

- why log?
 - (1) simplifies the subsequent mathematical analysis (2) good for underflow the numerical precision of the computer
- **sample mean:** maximizint (1.54) whith respect to μ

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n \quad (1.55)$$

(ex.1.11)

$$\begin{aligned} \frac{\partial}{\partial \mu} \left(\sum_{n=1}^N (x_n - \mu)^2 \right) &= 0 \\ \frac{\partial}{\partial \mu} \left(\sum_{n=1}^N (x_n^2 - 2x_n\mu + \mu^2) \right) &= 0 \\ \sum_{n=1}^N (-2x_n + 2\mu) &= 0 \\ \sum_{n=1}^N 2x_n &= 2N\mu \\ \therefore \mu_{ML} &= \frac{1}{N} \sum_{n=1}^N x_n \end{aligned}$$

- **sample variance:** maximizint (1.54) whith respect to σ^2
 - the solution (μ_{ML}) and σ_{ML}^2 is decoupled. \rightarrow calculation order does not matter

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2 \quad (1.55)$$

(ex.1.11)

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 \right) &= 0 \\ -\frac{1}{2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \sigma^2 &= 0 \\ \therefore \sigma_{ML}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2 \end{aligned}$$

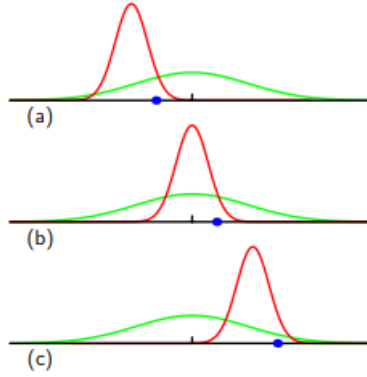
- **Limit: bias problem**

- μ_{ML} is unbiased, σ_{ML}^2 is biased
- at the variance, bias (underestimation) \rightarrow over fitting
- more complex models with many parameters \rightarrow more bias \rightarrow over fitting

$$\mathbb{E}[\mu_{ML}] = \mu \quad (1.57)$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{N-1}{N} \sigma^2 \quad (1.58)$$

$$\therefore \text{unbiased variable : } \tilde{\sigma}^2 = \frac{N}{N-1} \sigma^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{ML})^2 \quad (1.59)$$



1.2.5 Curve fitting re-visited

Goal: to predictions for the target variable t given some new value of the input variable x on the basis of a set of training data comprising N input values $x = (x_1, \dots, x_N)^T$ and their corresponding target values $t = (t_1, \dots, t_N)^T$ (from a probabilistic perspective)

assume that it is a Gaussian distribution

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (1.60)$$

Step 1: using the training set $\{\mathbf{x}, \mathbf{t}\} \rightarrow$ finding an unknown \mathbf{w} & β by maximum likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \quad (1.61)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \quad (1.62)$$

- \mathbf{w}_{ML} : minimize $\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$ (=1.2)
- β_{ML} :

$$\frac{1}{\beta} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{ML}) - t_n\}^2 \quad (1.62)$$

- Having determined the parameters \mathbf{w} and $\beta \rightarrow$ predictive distribution that gives the probability distribution over t

$$p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1}) \quad (1.63)$$

Step 2: introduce a prior distribution for Bayesian approach

- **prior**

$$p(\mathbf{w}, \alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\} \quad (1.65)$$

- α : precision (hyperparameter), $M + 1$: the total number of elements in the vector \mathbf{w} for an M_{th} order polynomial
- **posterior**

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}, \alpha) \quad (1.66)$$

- **MAP:** maximizing the posterior distribution, determine w by finding the most probable value of w given the data (a point estimate)

$$\text{minimize } \frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{ML}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} \quad (1.67)$$

- It is same as (1.4) with a regularization parameter given by $\lambda = \alpha/\beta$.

1.2.6 Bayesian curve fitting

- Marginalizations of $\mathbf{w} \rightarrow$ predict \mathbf{w} as a distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w} \quad (1.68)$$

- $p(t|x, \mathbf{w})$: (1.60), $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$: posterior, normalizing the right-hand side of (1.66)

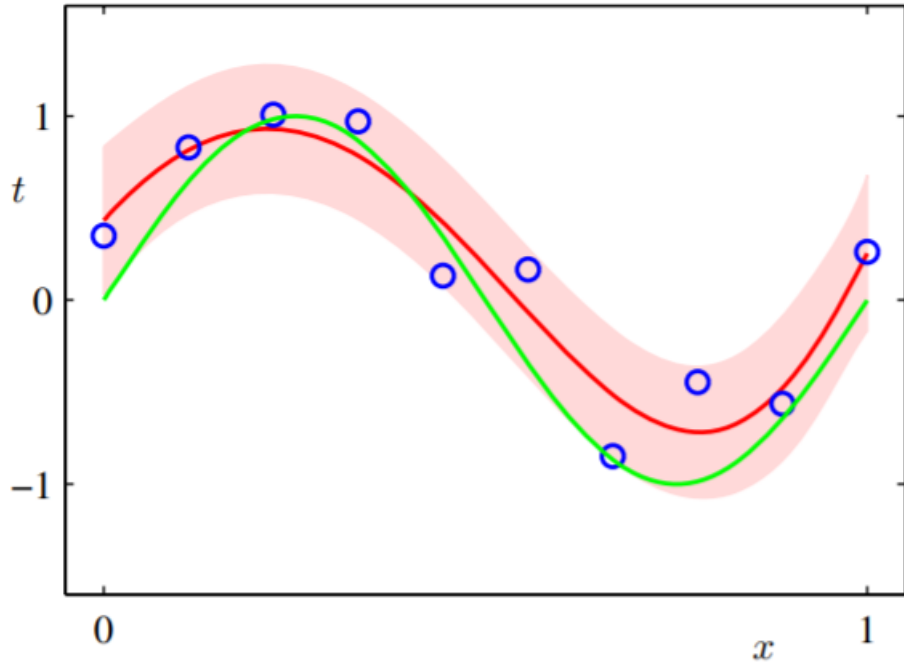
$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x)) \quad (1.69)$$

$$m(x) = \beta\phi(x)^T \mathbf{S} \sum_{n=1}^N \phi(x_n)t_n \quad (1.70)$$

$$s^2(x) = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x) \quad (1.70)$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(x_n)\phi(x_n)^T \quad (1.72)$$

- β^{-1} : the uncertainty in the predicted value of t , $\phi(x)^T \mathbf{S} \phi(x)$: the uncertainty of \mathbf{w} (a consequence of the Bayesian treatment)



1.3 Model Selection

- Select the number of free parameters (order)
- In the maximum likelihood approach, the performance on the training set is not a good indicator of predictive performance on unseen data (\therefore over-fitting)
- \therefore setting a validation set \rightarrow select the one having the best predictive performance
- However, the supply of data for training and testing will be limited \rightarrow **cross validation**
- **cross validation drawback:**
 - larger the number of factor of S, more the training runs
- **Information criteria**
 - akaike information criterion, AIC: add panalty term which is number of adjustable parameters at the log likelihood
 - Bayesian information criterion, BIC (Section 4.4.1)
- **Information criteria limits:**
 - not take account of the uncertainty in the model parameters
 - tend to favour overly simple model

1.4 The Curse of Dimensionality

- **The Curse of Dimensionality:** when the dimensionality increases, the volume of the space increases so fast that the available data become sparse

(example)

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^D w_i x_i + \sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j + \sum_{i=1}^D \sum_{j=1}^D \sum_{k=1}^D w_{ijk} x_i x_j x_k \quad (1.74)$$

- As the number of input variables D increases, so the number of independent coefficients $\propto D^3$
- For a polynomial of order M, the number of coefficients $\propto D^M$

1.5 Decision Theory

- When combined with probability theory, allows us to make optimal decisions in situations involving uncertainty
- **Inference:** Determination of $p(x, t)$ from a set of training data
 - $p(x, t)$: complete summary of the uncertainty associated with these variables
 - any of the quantities appearing in Bayes' theorem can be obtained from the joint distribution $p(x, t)$ by either marginalizing or conditioning with respect to the appropriate variables

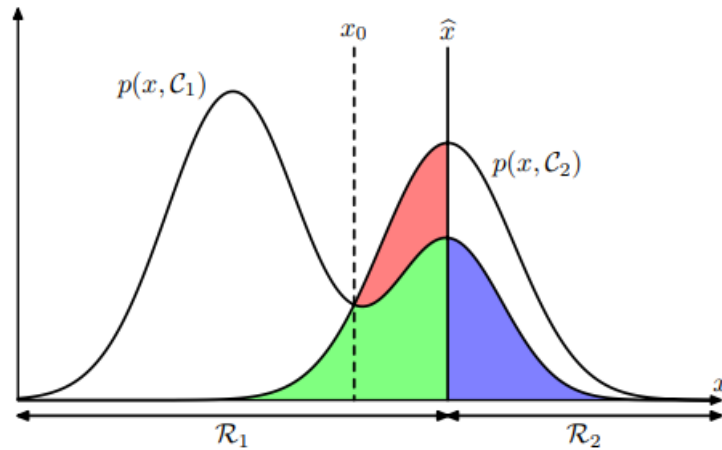
$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} \quad (1.77)$$

1.5.1 Minimizing the misclassification rate

Goal: to make as few misclassifications as possible

- Decision region: a rule - *assigns each value of x to one of the available classes*- will divide the input space into regions \mathcal{R}_k for each class, such that all points in \mathcal{R}_k are assigned to class C_k
- Decision boundary or decision surface: the boundaries between decision regions

$$\begin{aligned} \text{maximize } p(\text{correct}) &= \sum_{k=1}^K p(\mathbf{x} \in \mathcal{R}_k, C_k) \\ &= \sum_{k=1}^K \int_{\mathcal{R}_k} p(\mathbf{x}, C_k) d\mathbf{x} \end{aligned} \quad (1.79)$$



- The optimal choice for \hat{x} is where the curves for $p(x, C_1)$ and $p(x, C_2)$ cross, corresponding to $\hat{x} = x_0$, because in this case the red region disappears.

1.5.2 Minimizing the expected loss

- **cost function** or **loss function** : overall measure of loss incurred in taking any of the available decisions or actions, $L_{kj}p(\mathbf{x}, \mathcal{C}_k)$

Goal: to minimize the total loss incurred

(*loss matrix*)

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

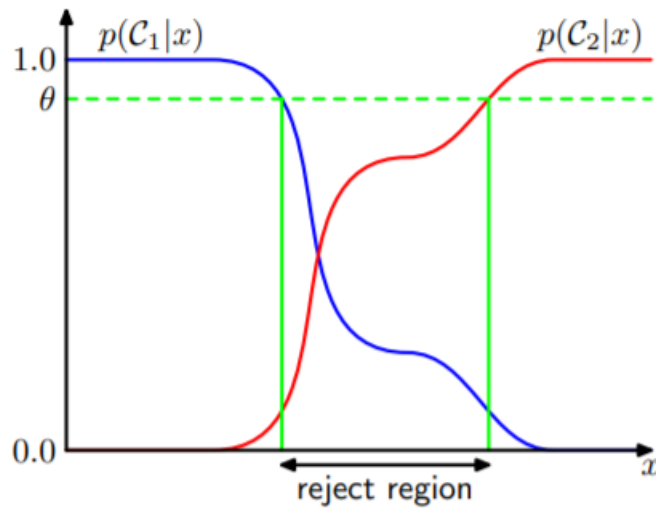
- The loss function depends on the true class, which is unknown.

$$\begin{aligned} \text{minimize } \mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \\ &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathcal{C}_k | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (1.80)$$

$$\text{minimize } \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \quad (1.81)$$

1.5.3 The reject option

- sometimes $p(\mathcal{C}_k | \mathbf{x})$ is too small (= joint distributions $p(x, \mathcal{C}_k)$ s are similar value)
- In areas where it is difficult to make a decision, the reject option could be better



1.5.4 Inference and decision

- Decision problem process: inference stage (train the posterior) \rightarrow decision stage *or* using discriminant function
- **(a) generative model**
 - (1) solve the inference problem, Determining the class-conditional densities $p(x|\mathcal{C}_k)$ for each class \mathcal{C}_k individually
 - (2) separately infer the prior class probabilities $p(C_k)$
 - or model the joint distribution $p(x, C_k)$ directly and then normalize
 - (3) obtain the posterior probabilities
 - Advantage: using $p(x) \rightarrow$ outlier detection or novelty detection
 - Limit: excessively demanding of data, to find the joint distribution
- **(b) discriminative model**
 - obtain a posterior probability directly
 - (1) solve the inference problem
 - (2) using the decision theory
 - (3) to assign each new \mathbf{x} to one of the classes
- **(c) using discriminant function**
 - discriminant function \rightarrow directly assigning
 - In this case, probabilities play no role
- The reasons for the posterior probabilities
 - Minimizing risk
 - Reject option
 - Compensating for class priors
 - Combining models

1.5.5 Loss functions for regression

Goal: to choose $y(x)$ so as to minimize the average, or expected, loss $\mathbb{E}[L]$.

$$\text{minimize } \mathbb{E}[L] = \int \int L(t, y(\mathbf{x})) p(\mathbf{x}, t) d\mathbf{x} dt \quad (1.86)$$

- A common loss function in regression problems: $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$
- **Regression function:** the conditional average of t conditioned on \mathbf{x}
 - The regression function $y(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$.

$$\text{minimize } \mathbb{E}[L] = \int \int \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt \quad (1.87)$$

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) dt \quad (1.88)$$

(appendix D)

$$\begin{aligned} \int y(\mathbf{x}) p(\mathbf{x}, t) dt - \int t p(\mathbf{x}, t) dt &= 0 \\ y(\mathbf{x}) p(\mathbf{x}) &= \int t p(\mathbf{x}, t) dt \\ y(x) &= \frac{\int t p(\mathbf{x}, t) dt}{p(x)} = \frac{\int t p(t|\mathbf{x}) p(x) dt}{p(x)} \\ y(x) &= \frac{\int t p(\mathbf{x}, t) dt}{p(x)} = \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}] \end{aligned} \quad (1.89)$$

- slightly different way,

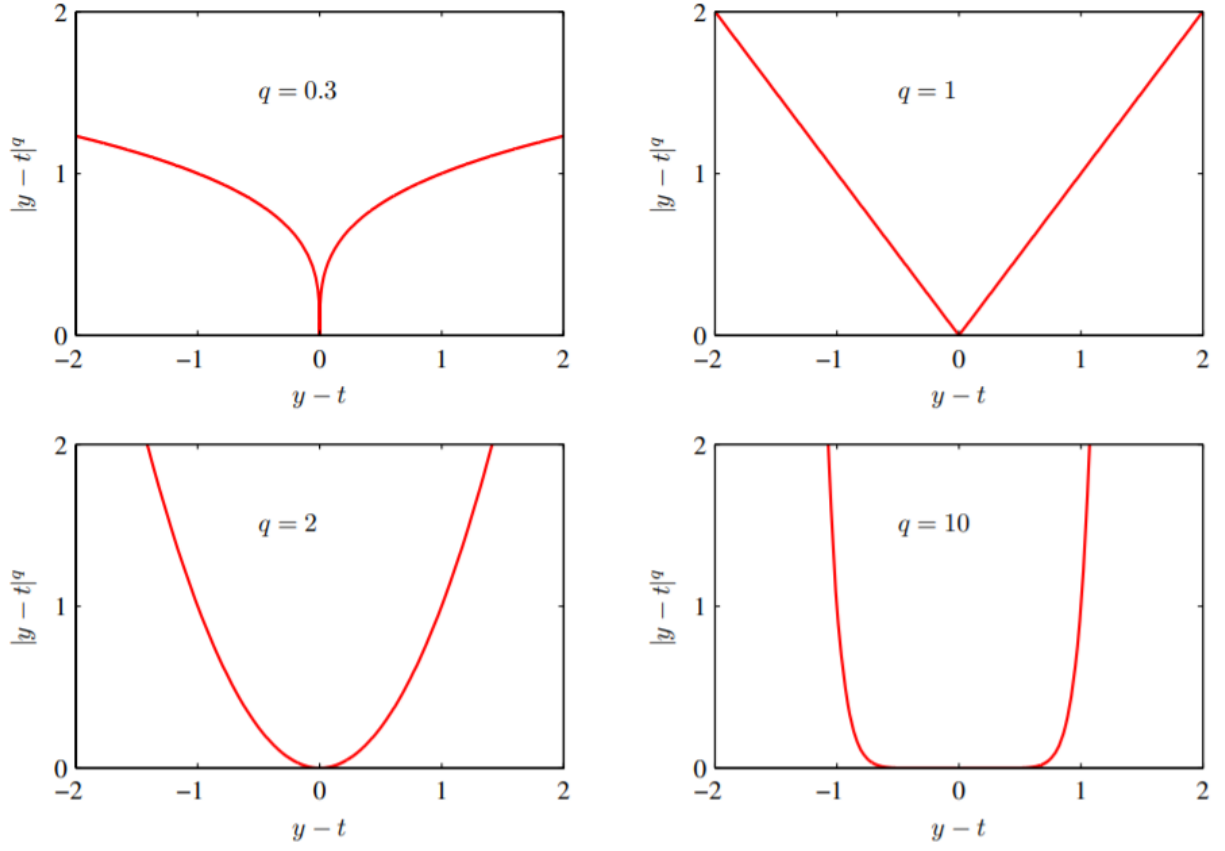
$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[L] &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}, t) dt \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}) d\mathbf{x} = \text{var}[t|\mathbf{x}] p(\mathbf{x}) \end{aligned} \quad (1.90)$$

- second term
 - the variance of the distribution of t , averaged over \mathbf{x}
 - the irreducible minimum value of the loss function, noise

- another loss function: Minkowski loss

$$\mathbb{E}[L_q] = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt \quad (1.91)$$



1.6 Information Theory

1.6.1 Relative entropy and mutual information