

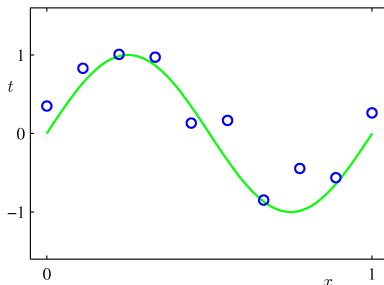
Introduction

Chapter 1

Prof. Reza Azadeh

University of Massachusetts Lowell

Polynomial Curve Fitting - An Example (1)



- We are given a training set comprising N observations of x , denoted as $(x_1, \dots, x_N)^\top$, together with corresponding observations of the values of t , denoted $(t_1, \dots, t_N)^\top$
- The data for this example is generated from $t = \sin(2\pi x) + \mathcal{N}(0, 0.3)$.

Polynomial Curve Fitting - An Example (2)

- Our goal is to use this training set to predict value \hat{t} of the target variable for some new value \hat{x} of the input variable.
- A simple approach is to use *curve fitting* assuming a polynomial function of the form

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

- M is the order of the polynomial, and x_j denotes x raised to the power of j . The vector \mathbf{w} denotes the polynomial coefficients.

Polynomial Curve Fitting - An Example (3)

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

- The coefficients will be determined by fitting the polynomial to the training data.
- This can be done by minimizing an *error function* that measures the misfit between the function $y(x, \mathbf{w})$, for any given value of \mathbf{w} , and the training set data points.

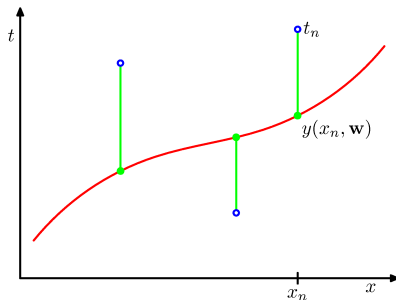
Polynomial Curve Fitting - An Example (4)

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

- One simple choice of the error function is *sum of the squares* of the errors between the predictions and the corresponding target values.
- So, we minimize a nonnegative quantity that would be zero if, and only if, the function $y(x, \mathbf{w})$ were to pass exactly through each training data point.

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

Polynomial Curve Fitting - An Example (5)



Sum-of-squares error function:

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

Polynomial Curve Fitting - An Example (6)

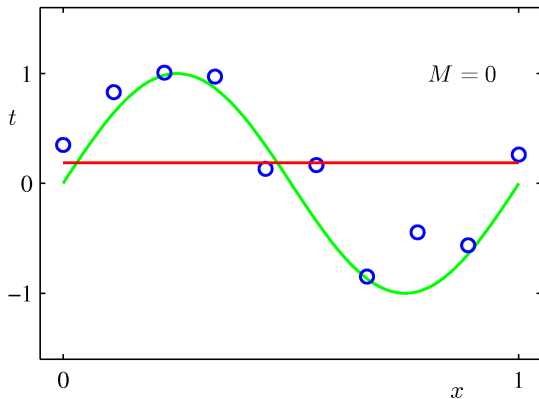
Because the error function is a quadratic function of the coefficients \mathbf{W} , its derivatives with respect to the coefficients will be linear in the elements of \mathbf{w} .

$$\begin{aligned}\mathbf{w}^* &= \text{minimize}_{\mathbf{w}} E(\mathbf{w}) \\ &= \text{minimize}_{\mathbf{w}} \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2\end{aligned}$$

For values of M , we show results of fitting polynomials having orders $M = 0, 1, 3$, and 9 .

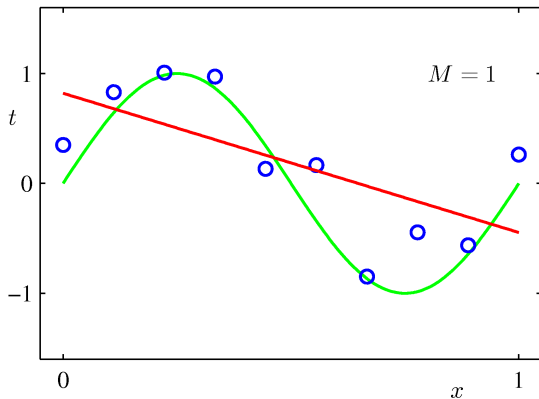
Polynomial Curve Fitting - An Example (7)

0th-order polynomial ($M = 0$)



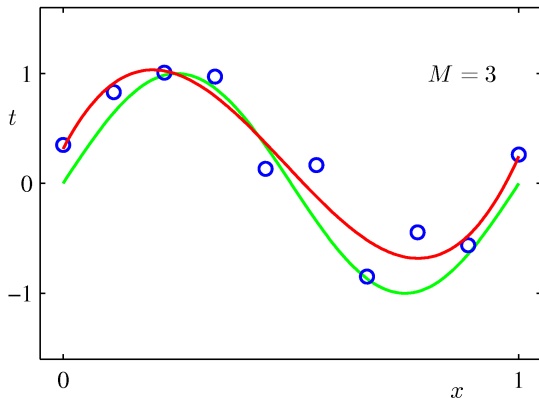
Polynomial Curve Fitting - An Example (8)

1st-order polynomial ($M = 1$)



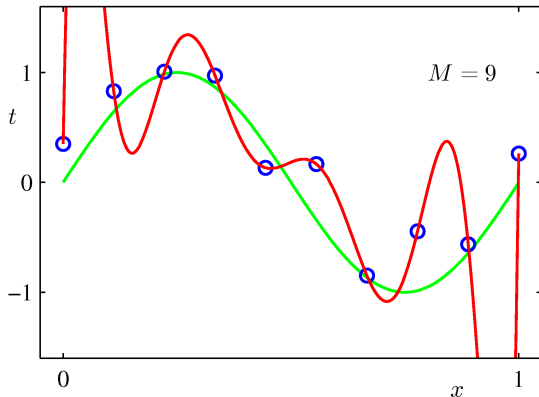
Polynomial Curve Fitting - An Example (9)

3rd-order polynomial ($M = 3$)



Polynomial Curve Fitting - An Example (10)

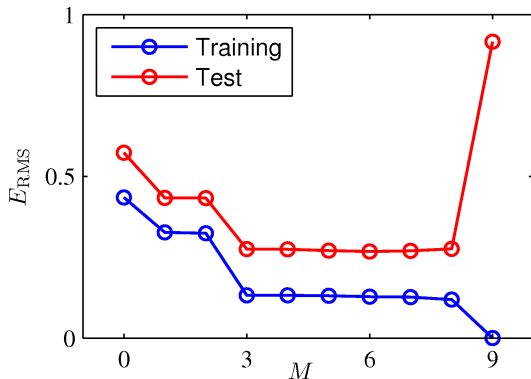
9th-order polynomial ($M = 9$)



\Rightarrow Over-fitting

Polynomial Curve Fitting - An Example (11)

Over-fitting



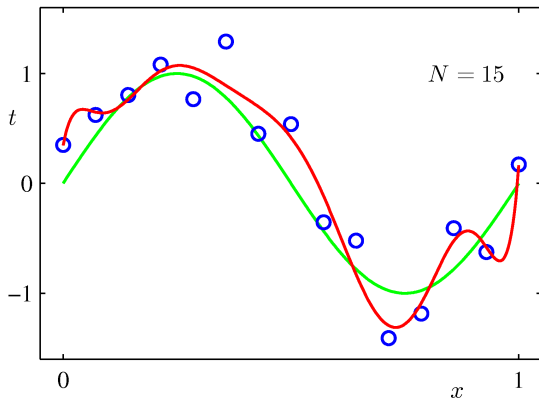
Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2(E(\mathbf{w}))/N}$

Polynomial Curve Fitting - An Example (12)

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
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

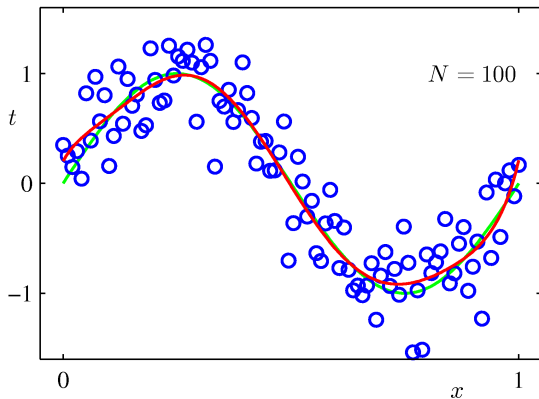
Polynomial Curve Fitting - An Example (13)

9th-order polynomial ($M = 9$) for data set size of 15.



Polynomial Curve Fitting - An Example (14)

9th-order polynomial ($M = 9$) for data set size of 100.



Polynomial Curve Fitting - An Example (15)

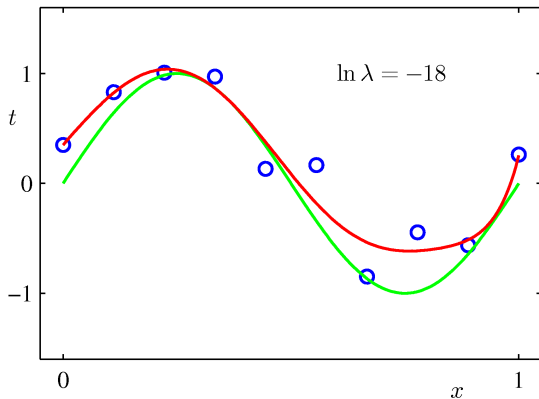
Regularization: adding a penalty term to the error function to discourage the coefficients from reaching large values.

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

where $\|\mathbf{w}\|^2 \equiv \mathbf{w}^\top \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$, and the coefficient λ governs the relative importance of regularization term compared with the sum-of-squares error term.

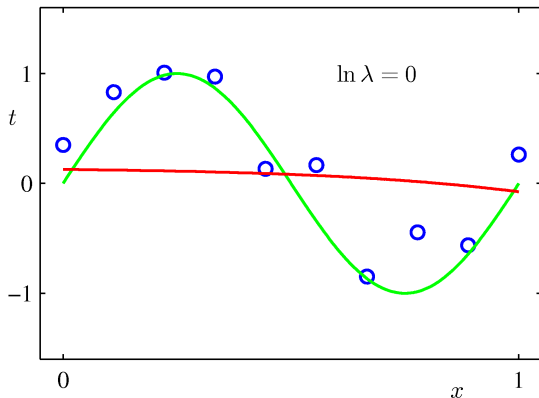
Polynomial Curve Fitting - An Example (16)

9th-order polynomial ($M = 9$) with $\ln \lambda = -18$.



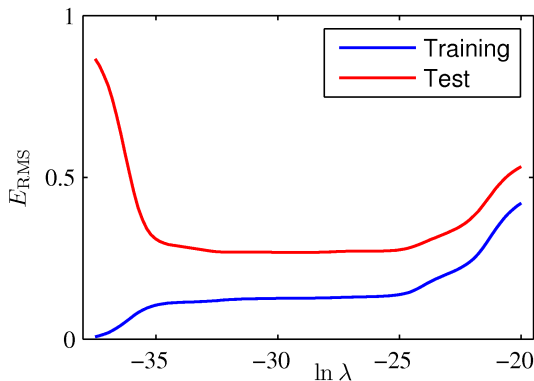
Polynomial Curve Fitting - An Example (17)

9th-order polynomial ($M = 9$) with $\ln \lambda = 0$.



Polynomial Curve Fitting - An Example (18)

E_{RMS} vs. $\ln \lambda$: in effect λ now controls the effective complexity of the model and hence determines the degree of over-fitting.



Polynomial Curve Fitting - An Example (19)

Polynomial coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
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

Probability Theory - Basic Concepts (1)

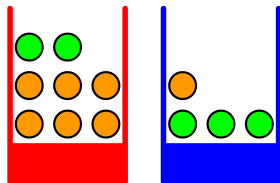
Random variables:

B (for boxes r and b),

F (for fruits a and o)

$$p(B = r) = 4/10$$

$$p(B = b) = 6/10$$

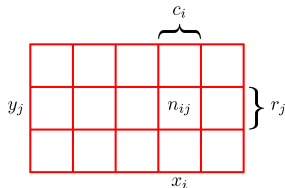


Probability Theory - Basic Concepts (2)

More generally, consider two random variables:

X (with values $\{x_i\}$, $i = 1, \dots, M$),

Y (with values $\{y_j\}$, $j = 1, \dots, L$),



A 3x5 grid of squares representing a contingency table. The columns are labeled x_i at the bottom, with a curly brace above the first four columns labeled c_i . The rows are labeled y_j on the left, with a curly brace to the right of the first two rows labeled r_j . The intersection of the second row and fourth column is labeled n_{ij} .

			n_{ij}	

- Consider a total of N trials in which we sample both of the variables X and Y . Let the number of trials in which $X = x_i$ and $Y = y_j$ be n_{ij} .
- Also let the number of trials in which X takes the value x_i irrespective of the value that Y takes be denoted as c_i . Similarly let the number of trials in which Y takes the value y_j be denoted by r_j .

Probability Theory - Basic Concepts (3)

Joint Probability:

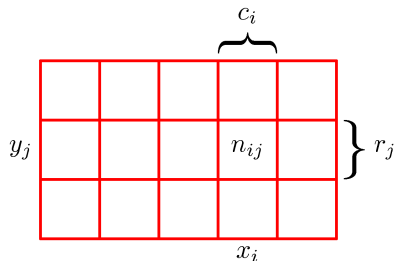
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal Probability:

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

Conditional Probability:

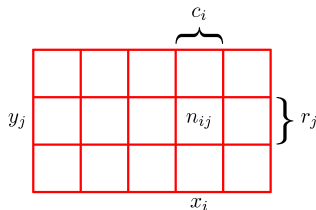
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



Probability Theory - Basic Concepts (4)

sum rule:

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} \\ &= \sum_{j=1}^L p(X = x_i, Y = y_j) \end{aligned}$$



product rule:

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

The Rules of Probability

- **sum rule**

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

- **product rule**

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

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

where $p(Y|X)$ is the *posterior* and $p(X) = \sum_Y p(X|Y)p(Y)$ is the normalization constant also known as the *evidence*.

Basic Concepts - Example (1)

Random variables:

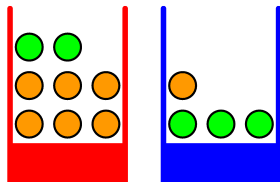
B (for boxes r and b),

F (for fruits a and o)

$$p(B = r) = 4/10$$

$$p(B = b) = 6/10$$

Note that these satisfy $p(B = r) + p(B = b) = 1$.

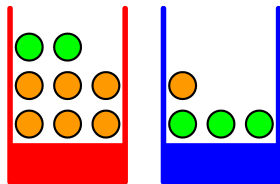


Basic Concepts - Example (2)

Suppose we pick a box randomly, and it turns out to be the blue box. Then the probability of selecting an apple is

$$p(F = a|B = b) = \frac{3}{4}$$

We can write all the conditional probabilities:



$$p(F = a|B = r) = \frac{1}{4}$$

$$p(F = o|B = r) = \frac{3}{4}$$

$$p(F = a|B = b) = \frac{3}{4}$$

$$p(F = o|B = b) = \frac{1}{4}$$

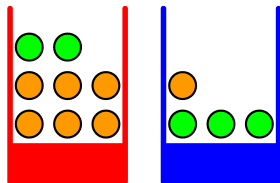
Basic Concepts - Example (3)

$$p(F = a|B = r) = \frac{1}{4}$$

$$p(F = o|B = r) = \frac{3}{4}$$

$$p(F = a|B = b) = \frac{3}{4}$$

$$p(F = o|B = b) = \frac{1}{4}$$



Note that the conditional probabilities are normalized so that

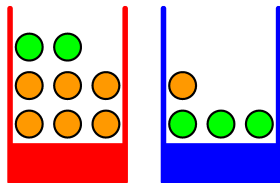
$$p(F = a|B = r) + p(F = o|B = r) = 1$$

$$p(F = a|B = b) + p(F = o|B = b) = 1$$

Basic Concepts - Example (4)

We can use the rules of probability to find the probability of choosing an apple:

$$p(F = a) = ?$$



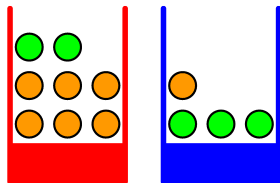
$$\begin{aligned} p(F = a) &= p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b) \\ &= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \end{aligned}$$

And then using the sum rule we get

$$p(F = o) = 1 - \frac{11}{20} = \frac{9}{20}$$

Basic Concepts - Example (5)

If a piece of fruit has been selected and it is an orange, calculate which box it came from.



We need to use the Bayes' theorem

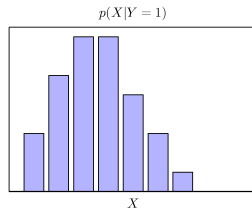
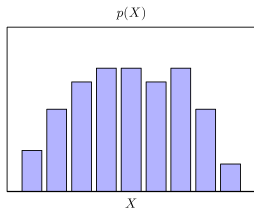
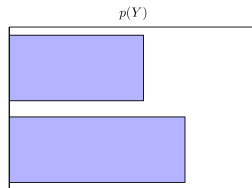
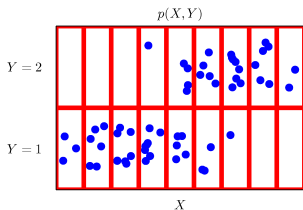
$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{(3/4) \times (4/10)}{9/20} = \frac{2}{3}$$

From the sum rule we then get

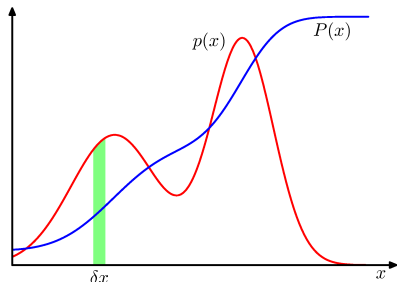
$$p(B = b|F = o) = 1 - \frac{2}{3} = \frac{1}{3}$$

Basic Concepts - Illustration

$N = 60$ samples for two random variables X (with 9 values) and Y (with 2 values).



Probability Densities (1)

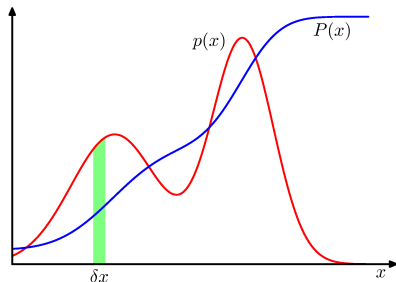


$$p(x \in (a, b)) = \int_a^b p(x) dx$$

Probability density $p(x)$ must satisfy the two conditions

$$p(x) \geq 0,$$
$$\int_{-\infty}^{\infty} p(x) dx = 1.$$

Probability Densities (2)



The probability that x lies in the interval $(-\infty, z)$ is given by the *Cumulative Distribution Function (CDF)* defined by

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

which satisfies $P'(x) = p(x)$.

Expectation

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

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

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

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

Conditional Expectation
(discrete)

Approximate Expectation
(discrete and continuous)

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

For the variable x itself, we can write

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

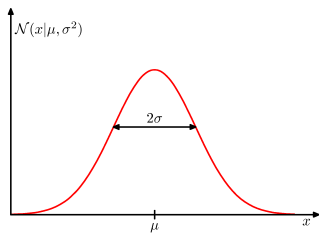
Note: in this course, we sometimes use $\mathbb{V}[x]$ instead of $\text{var}[x]$.

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

If x and y are independent then their covariance vanishes.
For two vectors of random variables, the covariance is a matrix

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

The Gaussian distribution



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

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

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

The Gaussian distribution - Mean and Variance

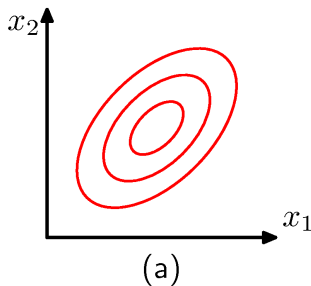
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu$$

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

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

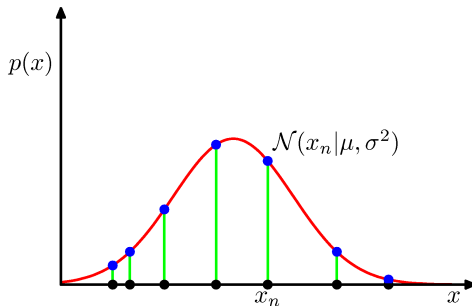
The Multivariate Gaussian

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



Gaussian Parameter Estimation

Likelihood function



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

Maximum (log) Likelihood

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

Maximizing with respect to μ ,

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

Maximizing with respect to σ^2 ,

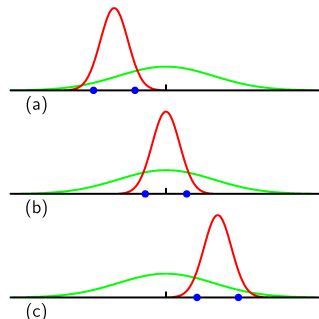
$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 \quad (\text{sample variance})$$

Properties of μ_{ML} and σ_{ML}^2 (1)

The maximum likelihood solutions are functions of the dataset values x_1, \dots, x_N . Consider the expectations of these quantities w.r.t the data set values. It can be shown that

$$\mathbb{E}[\mu_{\text{ML}}] = \mu$$

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



On average the maximum likelihood estimate will obtain the correct mean but will underestimate the true variance by a factor of $(N-1)/N$.

Properties of μ_{ML} and σ_{ML}^2 (2)

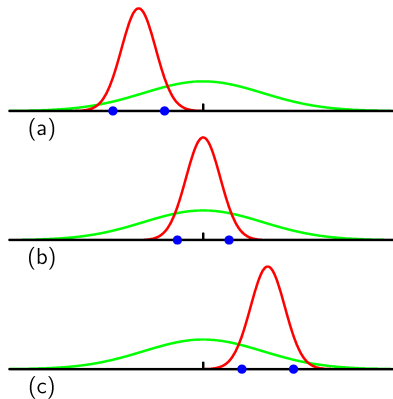
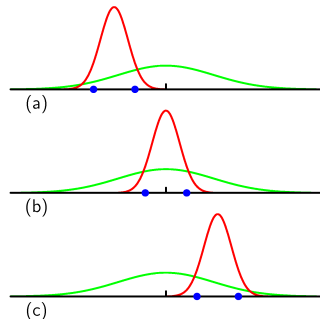


Figure: (green) true Gaussian, (red) estimated Gaussian from the data set (2 points)

Properties of μ_{ML} and σ_{ML}^2 (3)

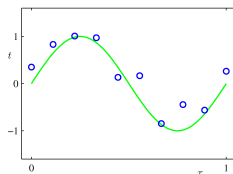
The following estimate for the variance parameter is unbiased

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{N}{N-1} \sigma_{\text{ML}}^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2\end{aligned}$$



Note that the bias of the maximum likelihood solution becomes less significant as the number of data points increases, and in the limit it becomes the true variance of the distribution.

Curve fitting re-visited (1)



- We are given a training set comprising N observations of x , denoted as $(x_1, \dots, x_N)^\top$, together with corresponding observations of the values of t , denoted $(t_1, \dots, t_N)^\top$.
- We can express our uncertainty over the value of the target variable using a probability distribution.
- We assume given x , the corresponding target t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ of the polynomial curve.

Curve fitting re-visited (2)

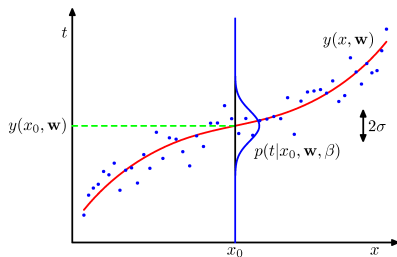
So we can write:

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

For the whole dataset, we have

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

Note: In this section, vectors \mathbf{t} and \mathbf{x} represent our 1-D data set.



Maximum Likelihood

We can make the log likelihood function as:

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

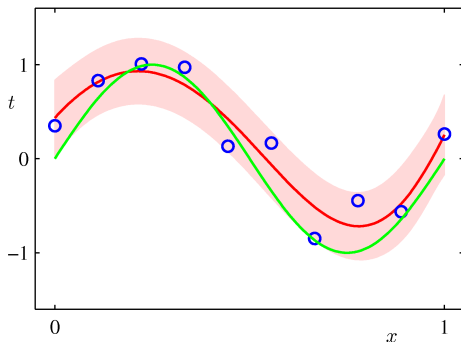
We can apply maximum likelihood to the first term given that the last two parts are constants. We also notice that the first term is the negative of $E(\mathbf{w})$.

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

Predictive Distribution

Having \mathbf{w}_{ML} and β_{ML} , we can now make predictions for new values of x .

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



MAP: A step toward Bayes (1)

For a more Bayesian approach, we can introduce a prior distribution over the polynomial coefficients \mathbf{w} . For simplicity, we consider a Gaussian distribution:

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

where α is the precision of the distribution, and $M + 1$ is the total number of elements in \mathbf{w} for an M^{th} order polynomial.

MAP: A step toward Bayes (2)

Now we can use the Bayes' theorem to form the posterior

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

We can determine \mathbf{w} by finding the most probable value of \mathbf{w} given the data, in other words by maximizing the posterior distribution. This is called **Maximum Posterior (MAP)**. In this case, we see that the maximum of the posterior is given by the minimum of

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

MAP: A step toward Bayes (3)

- Maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function with a regularization parameter given by $\lambda = \alpha/\beta$.
- \mathbf{w}_{MAP} is determined by minimizing the sum-of-squares errors, $\tilde{E}(\mathbf{w})$.

Bayesian curve fitting (1)

So far we have been making a point estimate of \mathbf{w} . In a fully Bayesian approach, we should consistently apply the sum and the product rules to integrate over all values of \mathbf{w} .

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

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

Note that we have omitted the α and β parameters.

Bayesian curve fitting (2)

The integration can be performed analytically:

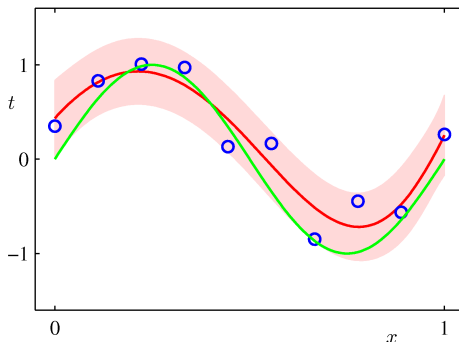
$$m(x) = \beta \phi(x)^\top \mathbf{S} \sum_{n=1}^N \phi(x_n) t_n$$
$$s^2(x) = \beta^{-1} + \phi(x)^\top \mathbf{S} \phi(x)$$

where

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(x_n) \phi(x_n)^\top$$
$$\phi(x_n) = (x_n^0, \dots, x_n^M)^\top$$

Bayesian curve fitting (3)

The predictive distribution resulting from a Bayesian treatment of polynomial curve fitting using $M = 9$ with $\alpha = 5 \times 10^{-3}$ and $\beta = 11.1$. (Red curve) mean, (Red area) ± 1 standard deviation around the mean.



We might want to train several models on the same data set and then find the best one for our application.

We already have seen that the performance of the training set is not a good indicator of predictive performance on unseen data due to overfitting.

What should we do?

Model selection when data is plentiful

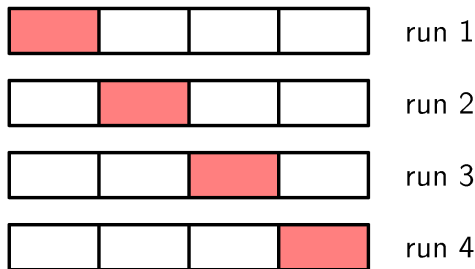
- If data is plentiful: use a portion of the data for training a range of models, then compare them on independent data, called *validation set*, and select the one that has the best predictive performance.
- If the model design is iterated many times using a limited size data set, then some overfitting to the validation data can occur and it may be necessary to keep aside a third *test set*.

Model selection when data is limited

- We would like to use as much of the available data as possible for training. However, if the validation set is small, it will give a relatively noisy estimate of predictive performance.
- solution: *cross-validation*

Cross Validation

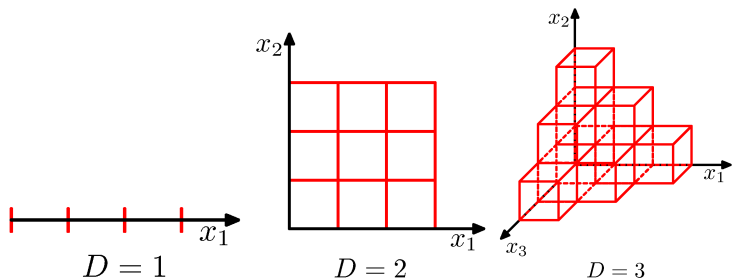
Benefit from limited training and testing data



S -Fold Cross Validation with ($S = 4$): partition the data into S groups. Use any $S - 1$ groups for training and the remainder for testing. The repeat for all S possible choices.

The Curse of Dimensionality (1)

What if there are more than one input variables (D)?



The Curse of Dimensionality (2)

The curve fitting example has one input variable ($D = 1$). If we have D input variables, a general polynomial with coefficients up to order 3 would take the form:

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

For a polynomial of order M , the growth in the number of coefficients is like D^M .

- **Inference step:**
determine either $p(t|\mathbf{x})$ or $p(t, \mathbf{x})$
- **Decision step:**
for given \mathbf{x} determine optimal t

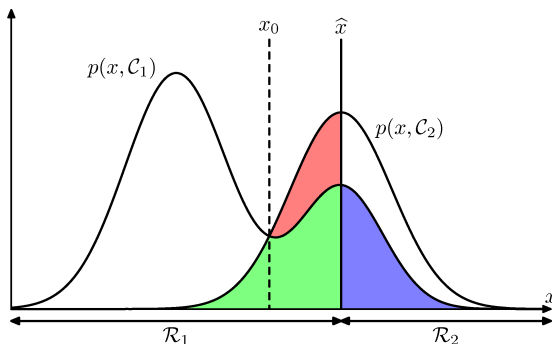
Given an image \mathbf{x} (i.e., the set of pixel intensities) determine if the patient has cancer (class \mathcal{C}_1 or $t = 0$) or not (class \mathcal{C}_2 or $t = 1$).

Inference step: The general inference problem determines the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ or $p(\mathbf{x}, \mathcal{C}_k)$, which gives us the most complete probabilistic description of the situation.

Decision step: In the end, we must decide either to give treatment to the patient or not, and we would like this choice to be optimal. In many cases, this can be solved very easily when we have solved the inference step,

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

Minimum Misclassification Rate



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

Minimum Expected Loss

An example of a loss matrix with elements L_{kj} for the cancer treatment problem. The rows correspond to the true class, whereas the columns correspond to the assignment of class made by our decision criterion.

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

Minimum Expected Loss

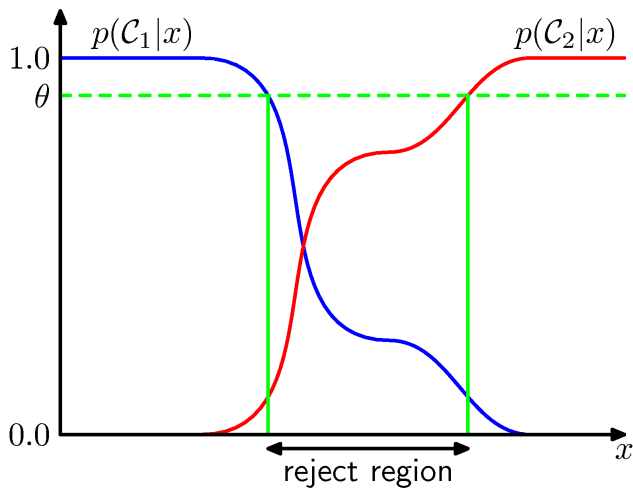
The expected loss can be written as

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

Regions \mathcal{R}_j are selected to minimize

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

Reject Option



Why separate inference and decision?

- minimizing risk (loss function might 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 \mathbf{x} , make optimal prediction $p(\mathbf{x})$, for t
- Loss function

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

The squared loss function

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

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

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

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

Generative vs. Discriminative

- **Generative approach**

model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

use Bayes' theorem $p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

- **Discriminative approach:**

model $p(t|\mathbf{x})$ directly

Consider a discrete random variable x and we ask how much information is received when we observe a specific value for this variable.

- If we are told that a **highly improbable** event has just happened, we will have received more information than if we were told that some **very likely** event has just occurred.
- If we know the event was certain to happen we would receive no information.

Information Theory (2)

Therefore, we look for a quantity $h(X)$ that expresses the information content and this should be a monotonic function of the probability $p(x)$.

- If we have two events x and y that are unrelated, then the information gain by observing both of them should be the sum of information gained from each of them separately so $h(x, y) = h(x) + h(y)$.
- Two unrelated events will be statistically independent and so $p(x, y) = p(x)p(y)$.
- from these two relationships, one can show that $h(x)$ must be given by the logarithm of $p(x)$ and so we have $h(x) = -\log_2 p(x)$.

$$H[x] = - \sum_x p(x) \log_2 p(x)$$

Important quantity in:

- coding theory
- statistical physics
- machine learning

Coding theory: consider x a discrete variable with 8 possible states; How many bit to transmit the state of x ? All states equally likely:

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

Now consider an example of a variable having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. Calculate the entropy:

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

average code length:

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

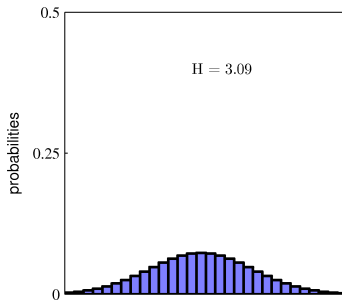
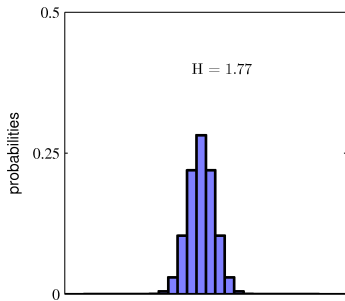
In how many ways can N identical objects be allocated M bins?

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

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

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

Entropy



Differential Entropy

Put bins of width Δ along the real line:

$$\lim_{\Delta \rightarrow 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) dx$$

Differential entropy maximized (for fixed σ^2) when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

in which case

$$H[x] = \frac{1}{2} \{ 1 + \ln(2\pi\sigma^2) \}$$

Conditional Entropy

$$H[\mathbf{y}|\mathbf{x}] = - \int \int p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{y} d\mathbf{x}$$

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$$

The Kullback-Leibler Divergence

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

$$\text{KL}(p\|q) \approx \frac{1}{N} \sum_{n=1}^N \{ -\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n) \}$$

$$\text{KL}(p\|q) \geq 0$$

$$\text{KL}(p\|q) \neq \text{KL}(q\|p)$$

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= \text{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y} \end{aligned}$$

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$$