機械学習 Machine Learning

序論 Introduction

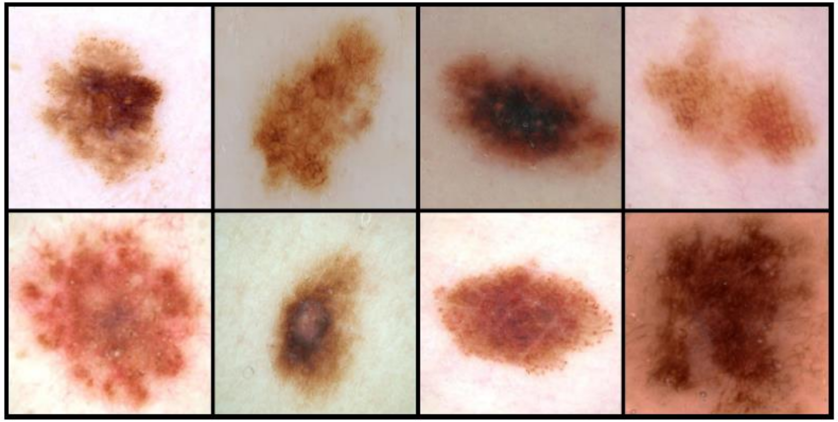
福嶋 誠 Makoto Fukushima

情報科学部
School of Informatics and Data Science

深層学習のインパクト The impact of deep learning
医療診断 Medical diagnosis

皮膚がん診断における問題 The problem of diagnosing skin cancer

メラノーマ（悪性黒色腫） Malignant melanomas	or	ほくろ（良性母斑） Benign nevi



深層学習のインパクト The impact of deep learning
タンパク質構造 Protein structure

タンパク質の3次元構造をそれらのアミノ酸配列から予測する.
Predict the 3D structures of proteins from their amino acid sequences.

Jumper et al. (2021)

深層学習のインパクト The impact of deep learning
画像生成 Image synthesis

ヒト顔画像を人工的に生成する.
Generate synthetic images of human faces.

https://generated.photos

深層学習のインパクト The impact of deep learning
大規模言語モデル Large language models

入力とした単語の系列における次の単語を生成する.
Generate a word that represents the next word in the input sentence.

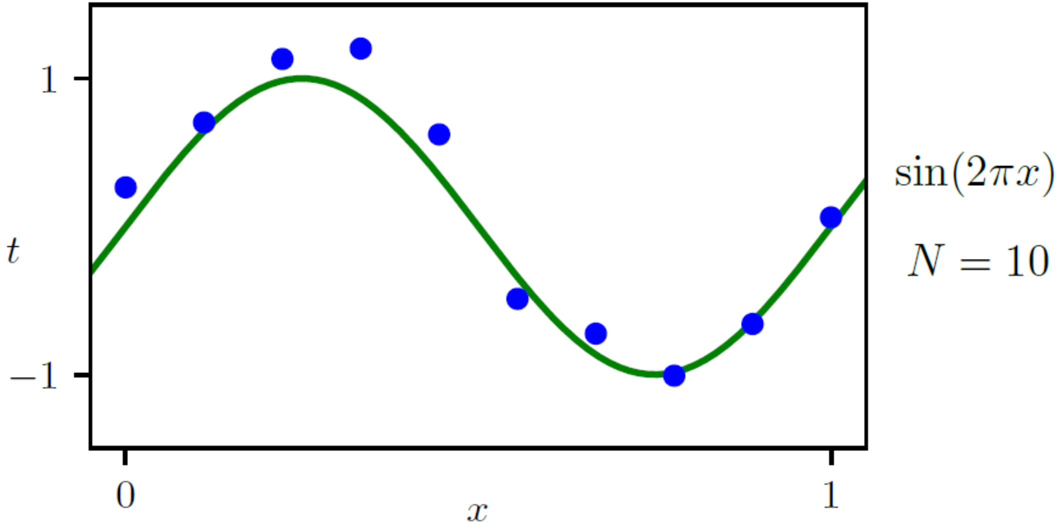
GPT-4: Proof of infinitely many primes

FIRST PARTY:
Thou dost not understand, my friend,
There simply is no upper end
To primes, those numbers indivisible,
No matter if it seems impossible.

SECOND PARTY:
But how dost thou so surely know?
To prove this truth, what canst thou show?

FIRST PARTY:
Consider this, my doubtful peer,
A clever proof that will make clear:
Assume that there’s a final prime,

チュートリアル例 A tutorial example
人工データ Synthetic data

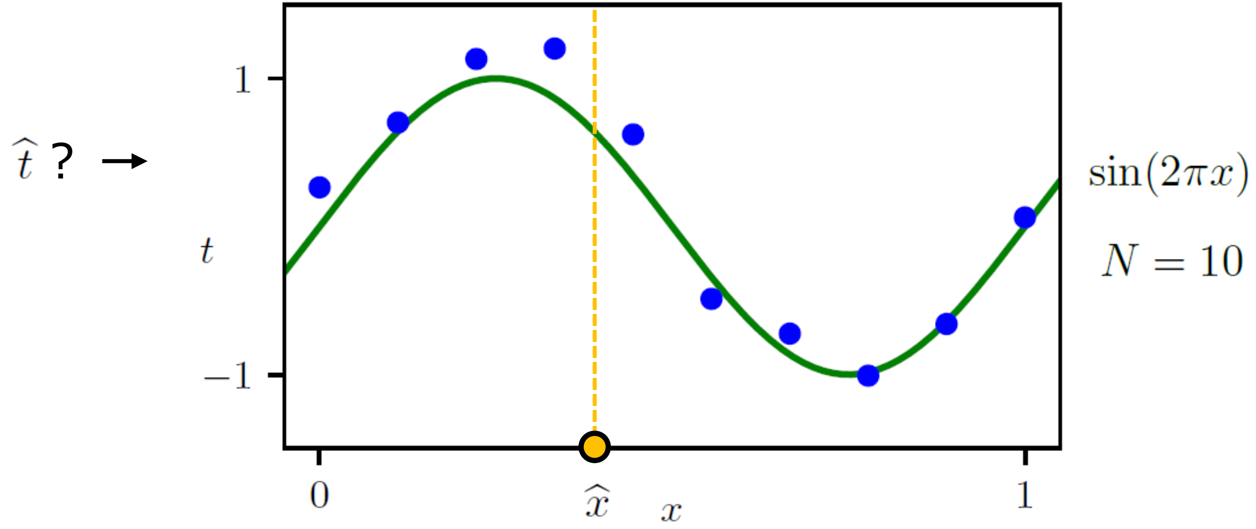


入力変数 Input variable: x — $x_1, \ldots, x_N$

目標変数 Target variable: t — $t_1, \ldots, t_N$

訓練集合
Training set

チュートリアル例 A tutorial example
人工データ Synthetic data

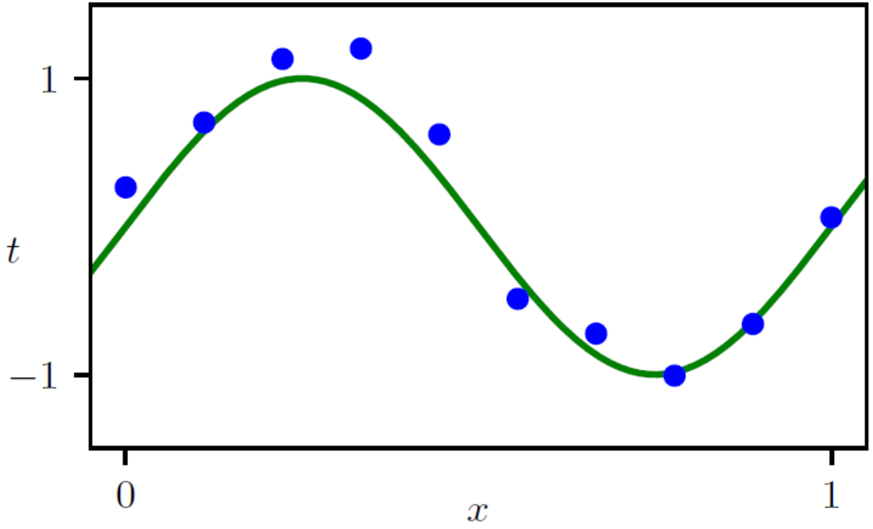


入力変数 Input variable: x $x_1, \ldots, x_N$

目標変数 Target variable: t $t_1, \ldots, t_N$

訓練集合
Training set

チュートリアル例 A tutorial example
線形モデル Linear models



多項式 Polynomial function

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

$\mathbf{w}$ の線形関数のひとつ A linear function of $\mathbf{w}$

チュートリアル例 A tutorial example
誤差関数 Error function

多項式 Polynomial function

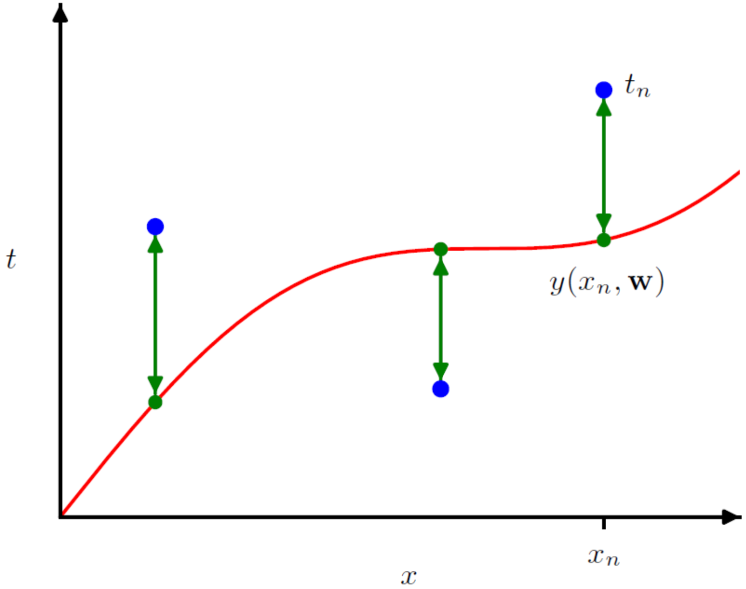
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

訓練集合 Training set

$$x_1, \ldots, x_N \qquad t_1, \ldots, t_N$$

誤差関数 Error function

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



チュートリアル例 A tutorial example
誤差関数 Error function

多項式 Polynomial function

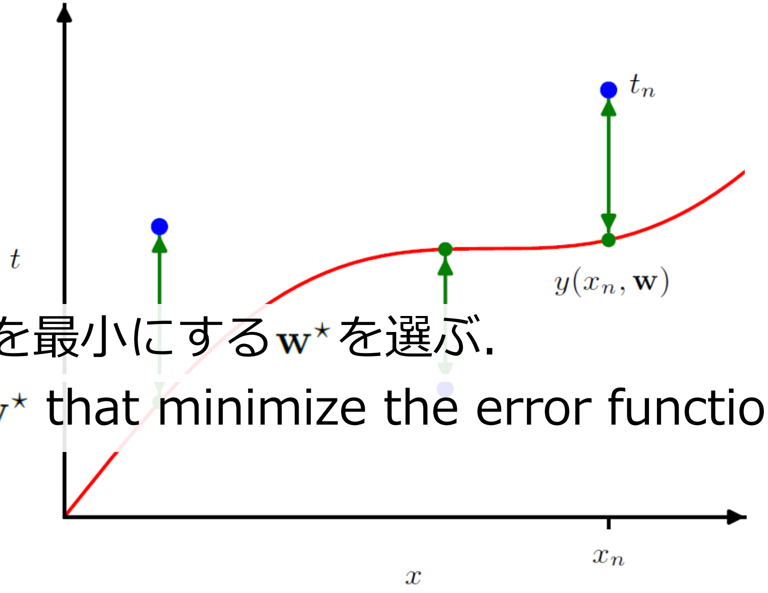
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

訓練集合 Training set

$$x_1, \ldots, x_N \quad t_1, \ldots, t_N$$

誤差関数 Error function

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

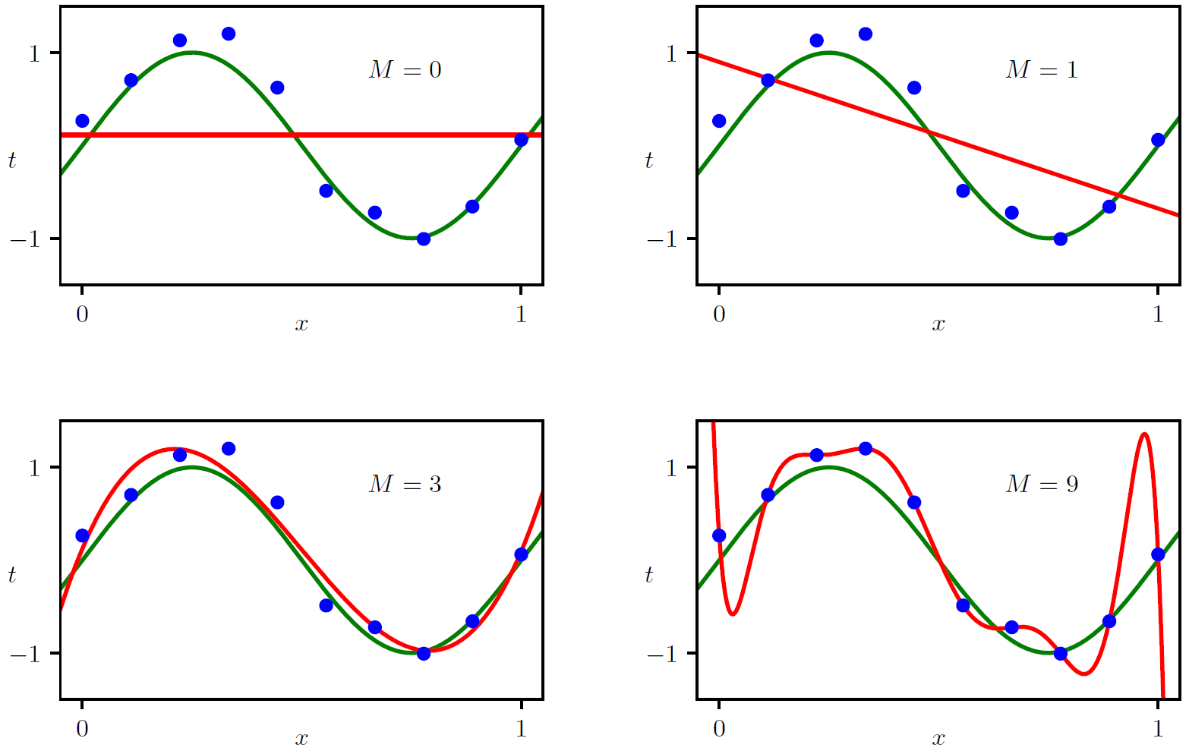


誤差関数を最小にする$\mathbf{w}^\star$を選ぶ.

Choose $\mathbf{w}^\star$ that minimize the error function.

チュートリアル例 A tutorial example
モデルの複雑さ Model complexity

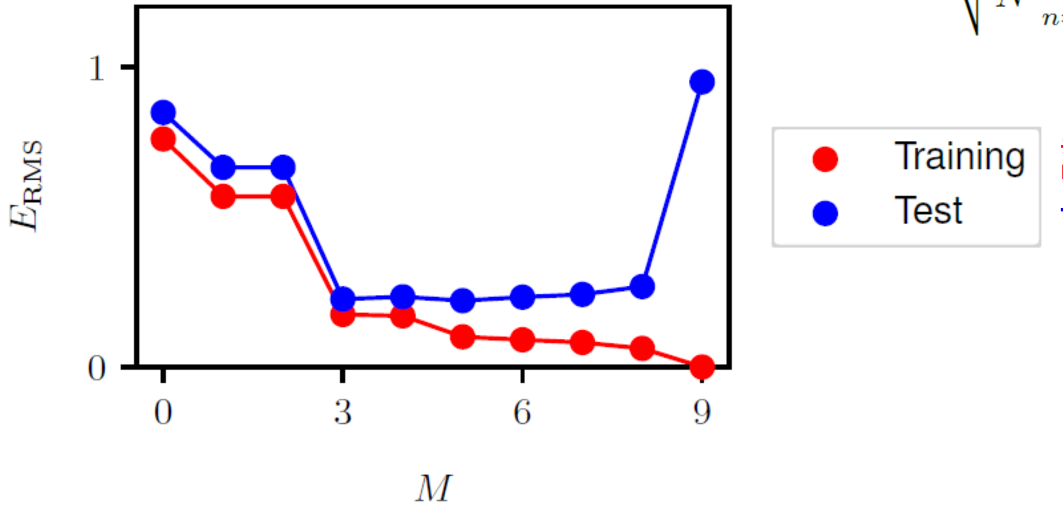
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



← 過学習
Over-fitting

チュートリアル例 A tutorial example
モデルの複雑さ Model complexity

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



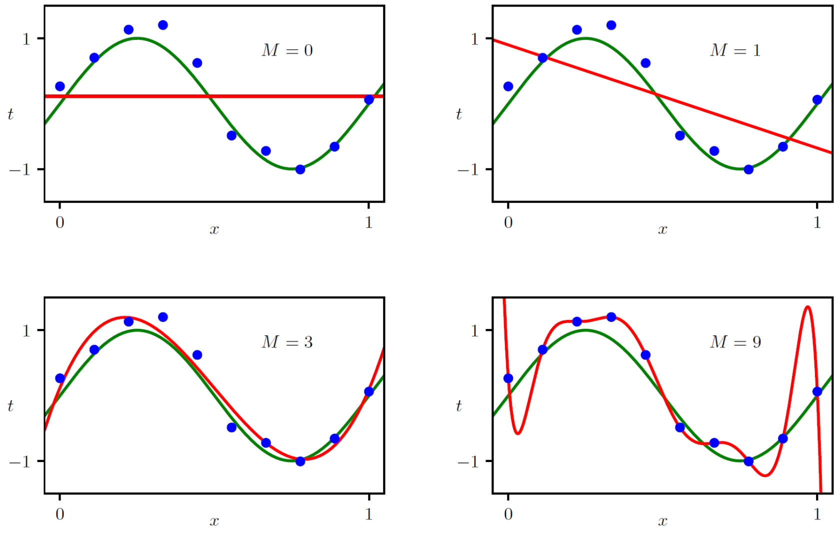
訓練
テスト

訓練集合を用いたときの誤差 Error derived from the training set

テスト集合を用いたときの誤差 Error derived from the test set

チュートリアル例 A tutorial example
モデルの複雑さ Model complexity

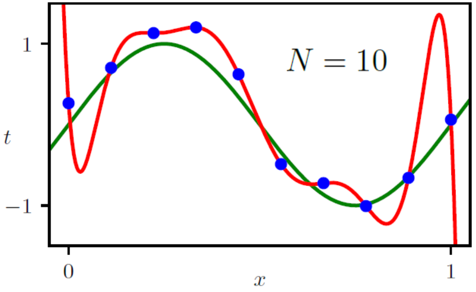
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

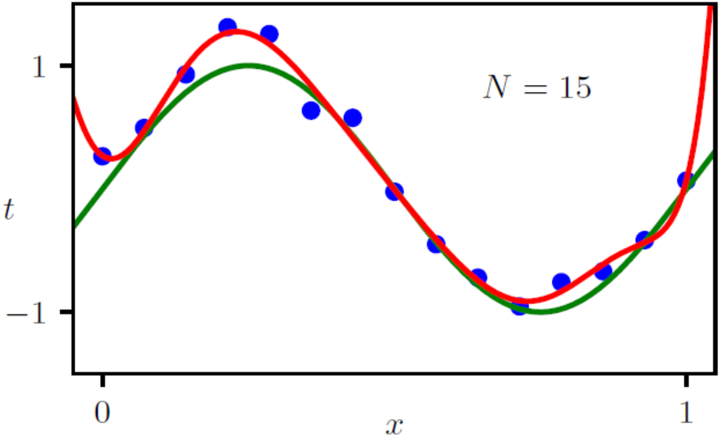


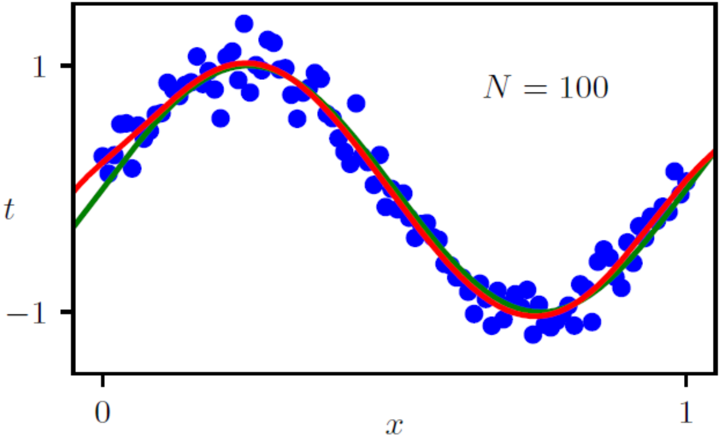
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.11	0.90	0.12	0.26
$w_1^\star$		−1.58	11.20	−66.13
$w_2^\star$			−33.67	1,665.69
$w_3^\star$			22.43	−15,566.61
$w_4^\star$				76,321.23
$w_5^\star$				−217,389.15
$w_6^\star$				370,626.48
$w_7^\star$				−372,051.47
$w_8^\star$				202,540.70
$w_9^\star$				−46,080.94

ランダムノイズに引きずられている
Increasingly tuned to the random noise

チュートリアル例 A tutorial example
モデルの複雑さ Model complexity







データ集合のサイズが大きくなるにつれて過学習の問題は深刻でなくなる.
The over-fitting problem become less severe as the size the data set increases.

チュートリアル例 A tutorial example
正則化 Regularization

正則化によりモデル複雑さをコントロールする.
Control the model complexity by regularization.

誤差関数に正則化項を付加する.
Add a regularization term to the error function.

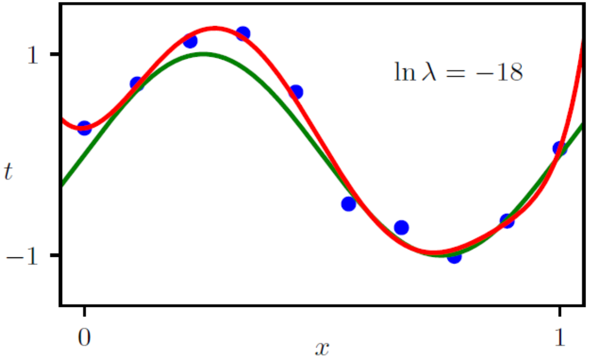
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \underline{\frac{\lambda}{2}\|\mathbf{w}\|^2}$$

正則化項
Regularization term

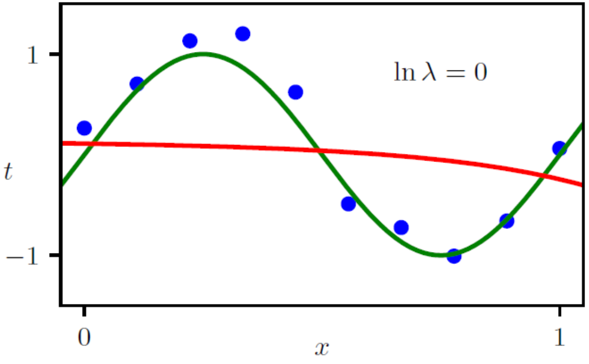
$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

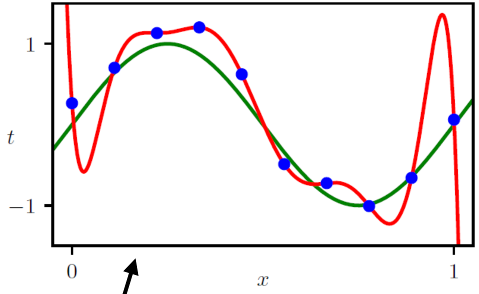
チュートリアル例 A tutorial example
正則化 Regularization

弱い正則化
Weak regularization



強い正則化
Strong regularization



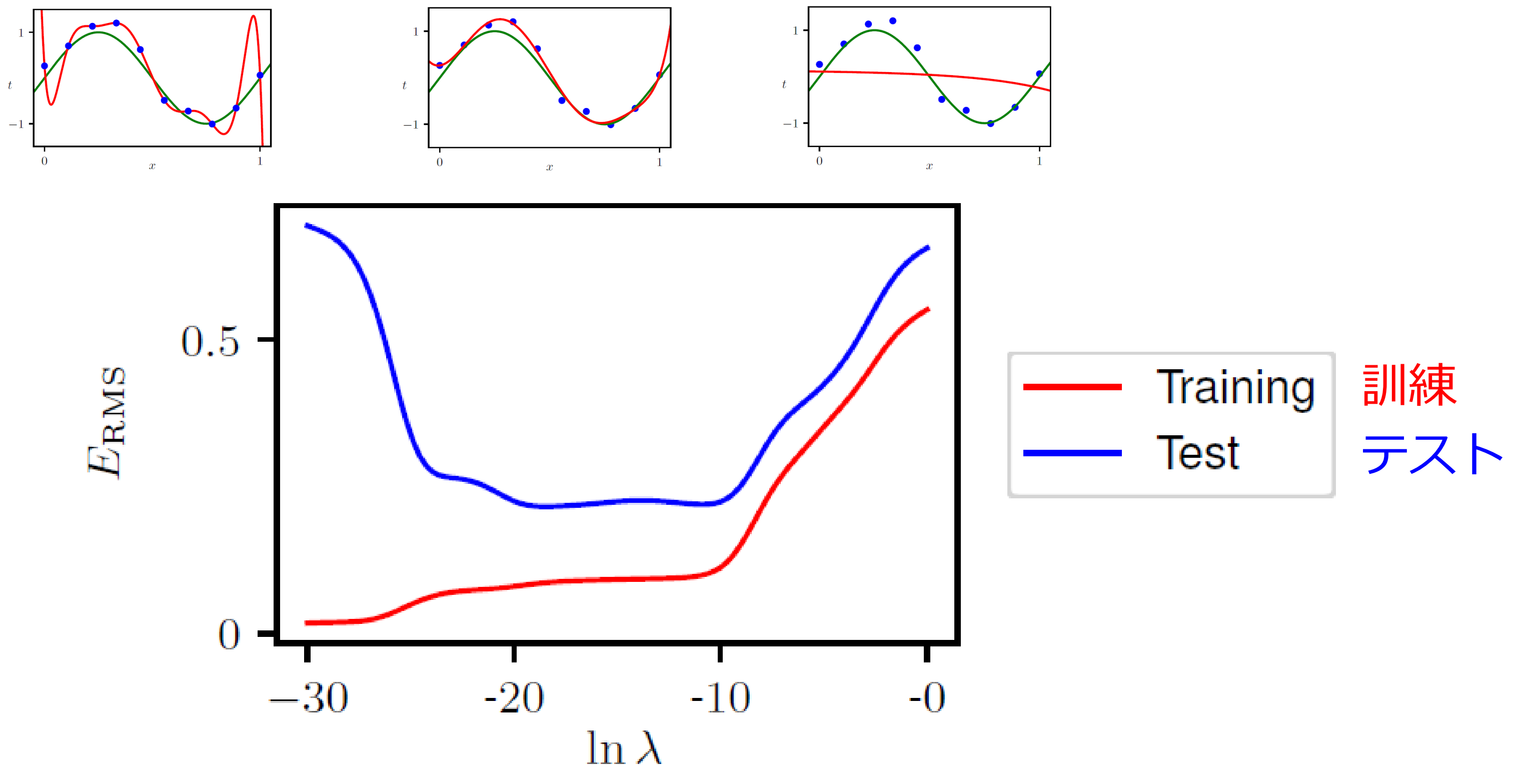


正則化なし No regularization

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.26	0.26	0.11
$w_1^\star$	−66.13	0.64	−0.07
$w_2^\star$	1,665.69	43.68	−0.09
$w_3^\star$	−15,566.61	−144.00	−0.07
$w_4^\star$	76,321.23	57.90	−0.05
$w_5^\star$	−217,389.15	117.36	−0.04
$w_6^\star$	370,626.48	9.87	−0.02
$w_7^\star$	−372,051.47	−90.02	−0.01
$w_8^\star$	202,540.70	−70.90	−0.01
$w_9^\star$	−46,080.94	75.26	0.00

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

チュートリアル例 A tutorial example
正則化 Regularization



チュートリアル例 A tutorial example
モデル選択 Model selection

<u>訓練集合 Training set</u>

パラメータ推定に用いる.
Use it for parameter estimation.

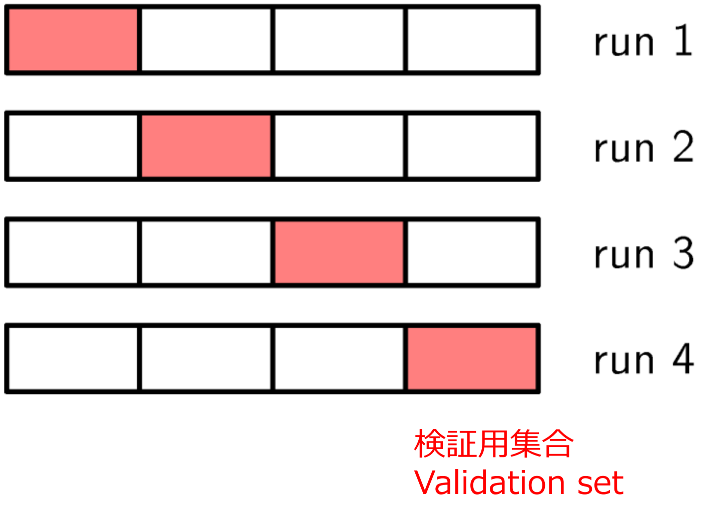
<u>検証用集合 Validation set</u>

モデル選択に用いる.
Use it for model selection.

<u>テスト集合 Test set</u>

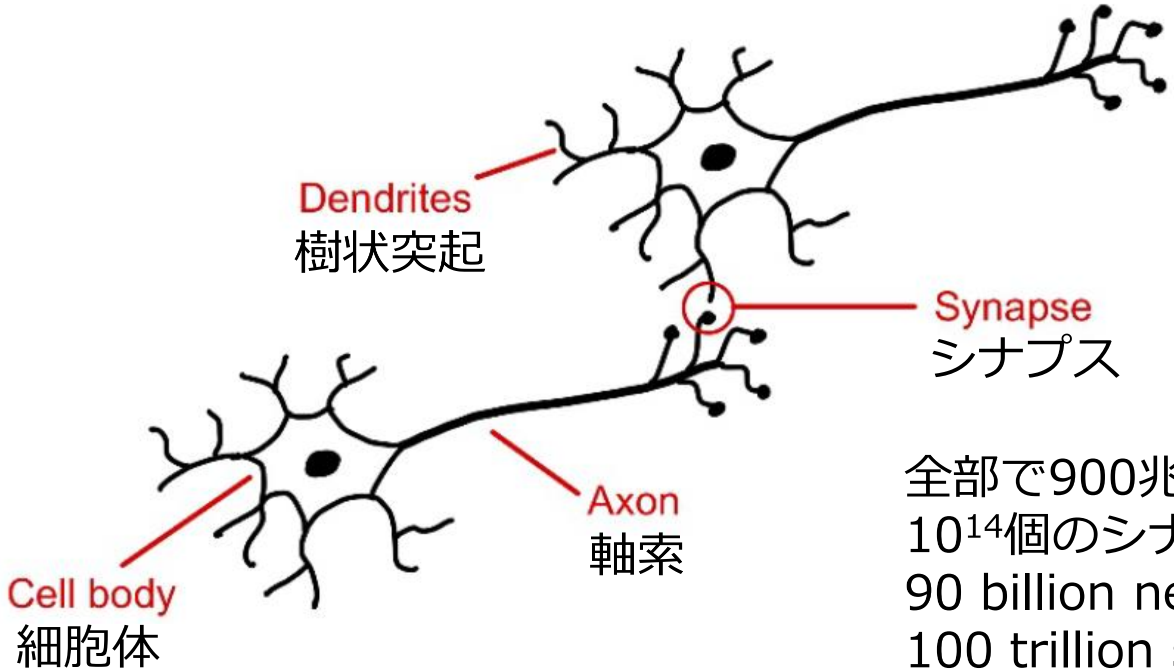
最終的な性能評価に用いる.
Use it for the final performance evaluation.

<u>交差検証 Cross-validation</u>



機械学習の歴史概要 A brief history of machine learning

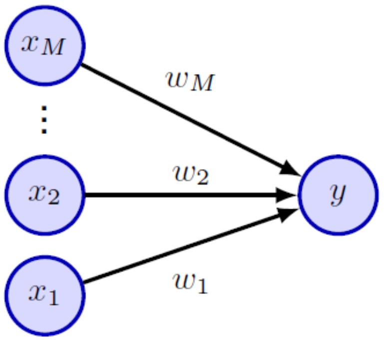
ヒト脳内における二個の神経細胞（ニューロン）の模式図
Schematic illustration showing two neurons from the human brain



全部で900兆個のニューロンと
10^{14}個のシナプス
90 billion neurons and
100 trillion synapses in total

機械学習の歴史概要 A brief history of machine learning

人工ニューラルネットワーク
Artificial neural networks McCulloch and Pitts (1943)



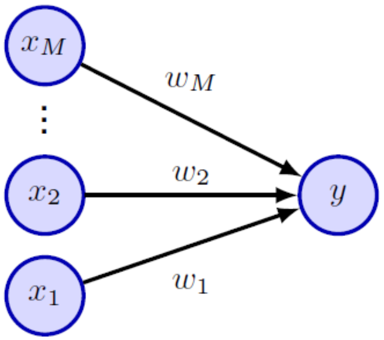
$$a = \sum_{i=1}^{M} w_i x_i$$

$$y = f(a)$$

$x_1, \ldots, x_M$	他のニューロンの活動 Activities of other neurons
$w_1, \ldots, w_M$	重み Weights
$f(\cdot)$	活性化関数 Activation function

機械学習の歴史概要 A brief history of machine learning

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



特殊なケース
A special case

$$a = \sum_{i=1}^{M} w_i x_i$$

$$y = f(a)$$

$x_1, \ldots, x_M$ 他のニューロンの活動 Activities of other neurons

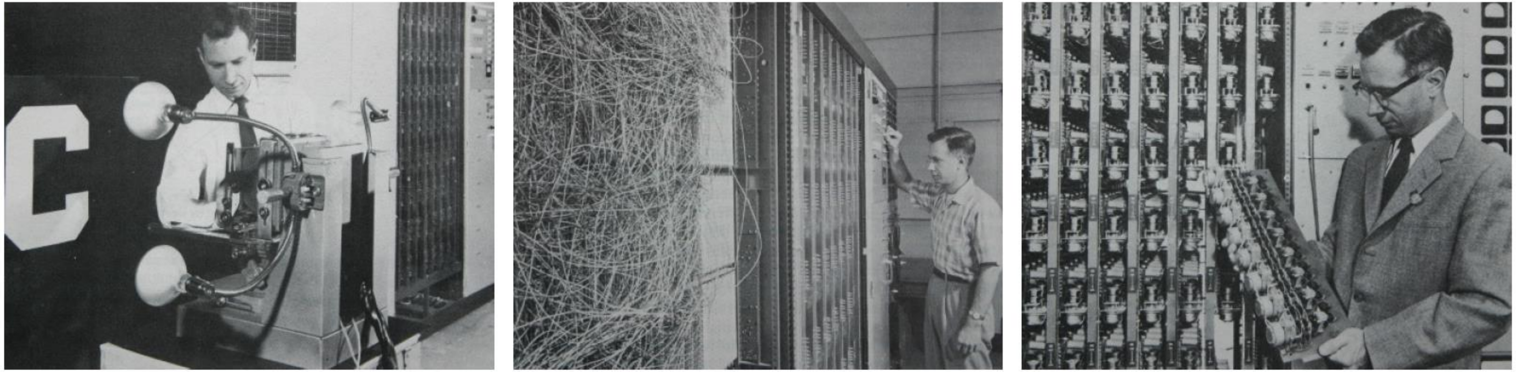
$w_1, \ldots, w_M$ 重み Weights

$f(\cdot)$ 活性化関数 Activation function

機械学習の歴史概要 A brief history of machine learning
単層ネットワーク Single-layer networks

パーセプトロン Perceptron Rosenblatt (1962)

$$f(a) = \begin{cases} 0, & \text{if } a \leqslant 0 \\ 1, & \text{if } a > 0 \end{cases}$$



Mark 1 perceptron hardware

機械学習の歴史概要 A brief history of machine learning
単層ネットワーク Single-layer networks

パーセプトロン Perceptron　Rosenblatt (1962)

$$f(a) = \begin{cases} 0, & \text{if } a \leqslant 0 \\ 1, & \text{if } a > 0 \end{cases}$$

訓練データ集合が線形分離可能である場合，パーセプトロンの学習アルゴリズムは有限回の繰り返しで厳密解に収束することが保証されている.

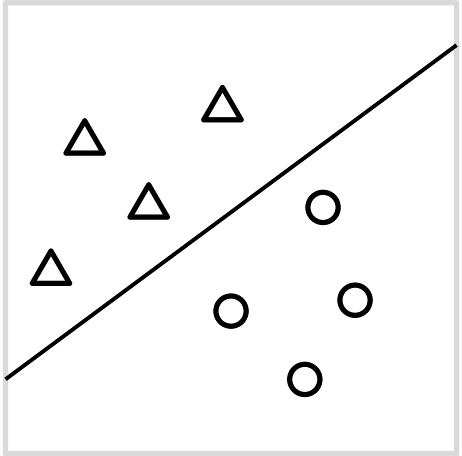
If the training data set is linearly separable, the perceptron learning algorithm is guaranteed to find an exact solution in a finite number of steps.

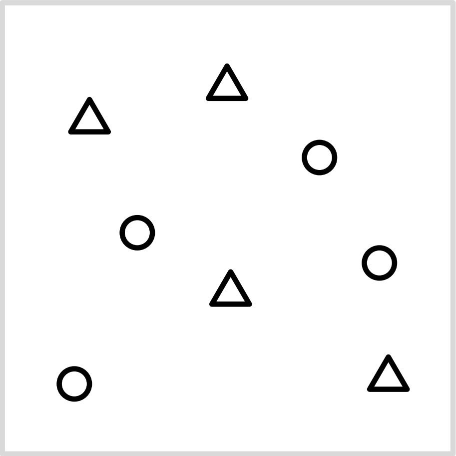
機械学習の歴史概要 A brief history of machine learning
単層ネットワーク Single-layer networks

パーセプトロン Perceptron　Rosenblatt (1962)

$$f(a) = \begin{cases} 0, & \text{if } a \leqslant 0 \\ 1, & \text{if } a > 0 \end{cases}$$

線形分離可能
Linearly separable

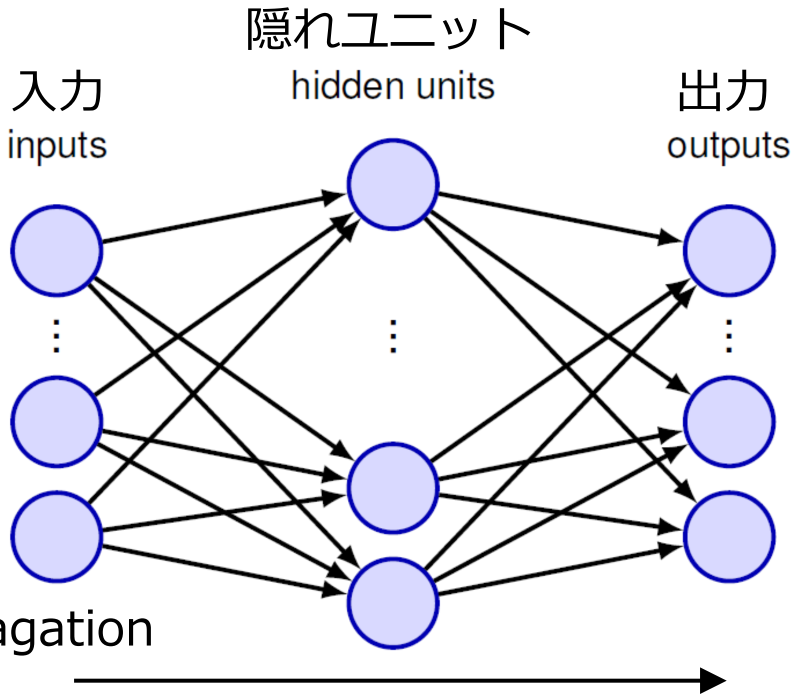




線形分離不可能
Not linearly separable

機械学習の歴史概要 A brief history of machine learning
逆伝播 Backpropagation

フィードフォーワードニューラルネットワーク
Feed-forward neural networks



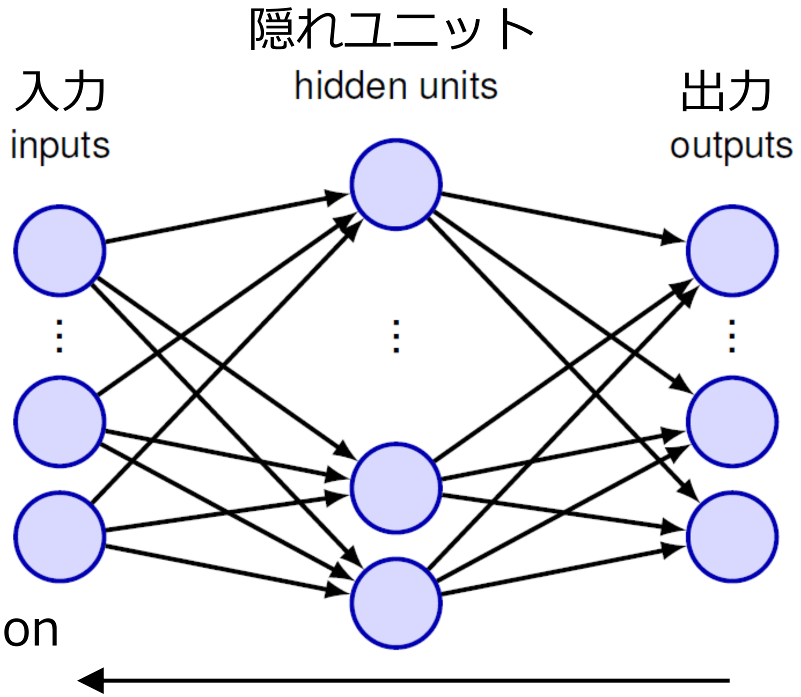
$$a = \sum_{i=1}^{M} w_i x_i$$

$$y = f(a)$$

順伝播 Forward propagation

機械学習の歴史概要 A brief history of machine learning
逆伝播 Backpropagation

フィードフォワードニューラルネットワーク
Feed-forward neural networks



$$a = \sum_{i=1}^{M} w_i x_i$$

$$y = f(a)$$

逆伝播 Backpropagation

Rumelhart et al. (1986)

機械学習の歴史概要 A brief history of machine learning
深層ネットワーク Deep networks

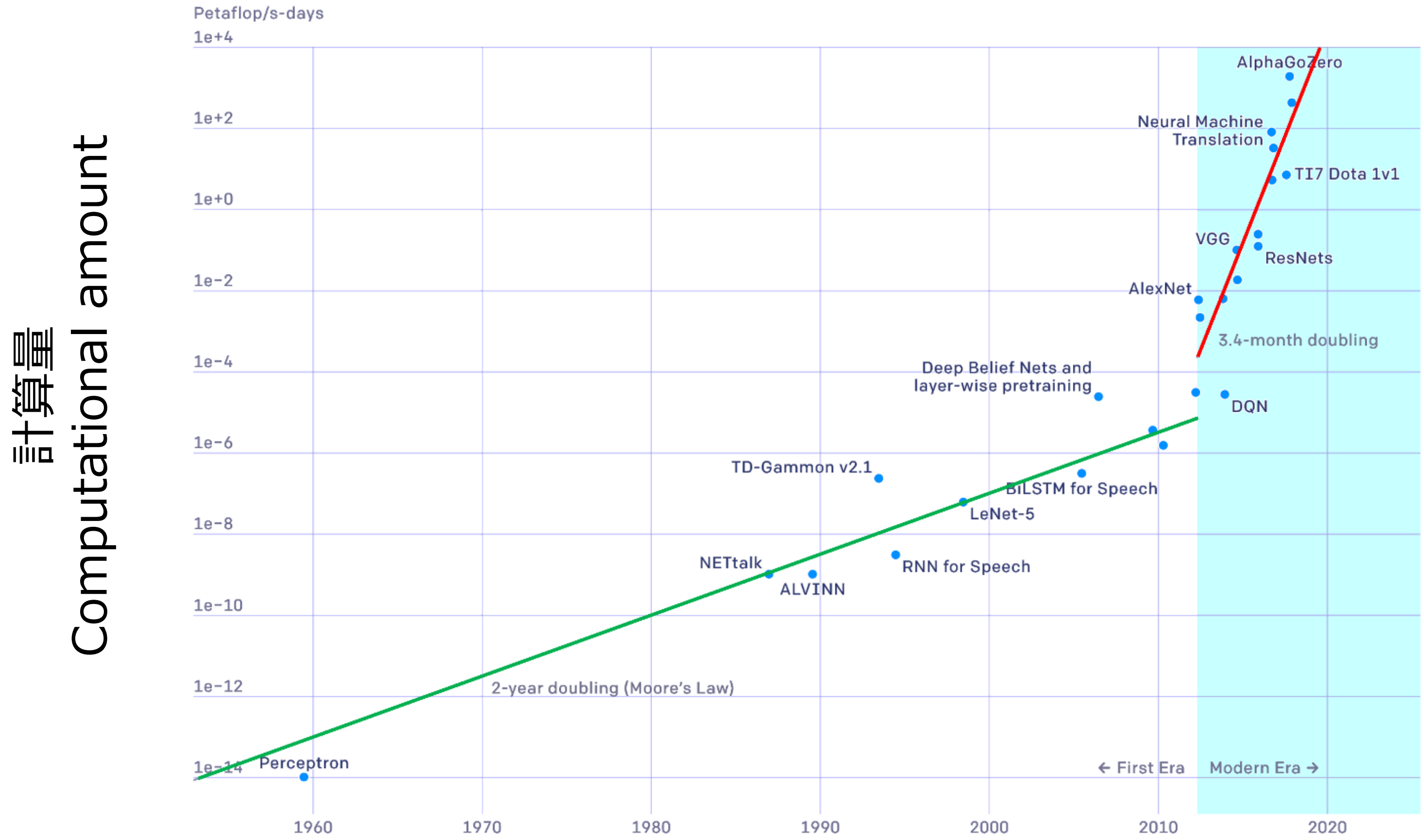
深層ニューラルネットワーク Deep neural networks

多数の重み層をもつニューラルネットワーク
Neural networks with many layers of weights

深層学習 Deep learning LeCun et al. (2015)

深層ニューラルネットワークに特化した機械学習の一領域
A sub-field of machine learning that focuses on deep neural networks

機械学習の歴史概要 A brief history of machine learning
深層ネットワーク Deep networks



From OpenAI

確率 Probabilities

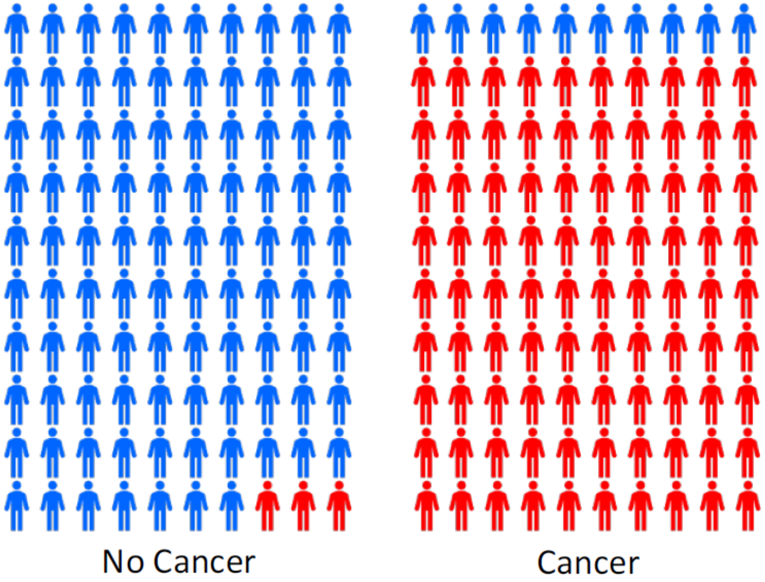
不確実性は確率論のフレームワークを用いて取り扱うことができる.
Uncertainty can be handled using the framework of probability theory.

60%

40%

確率の基本法則 The rules of probability
医療スクリーニング例 A medical screening example

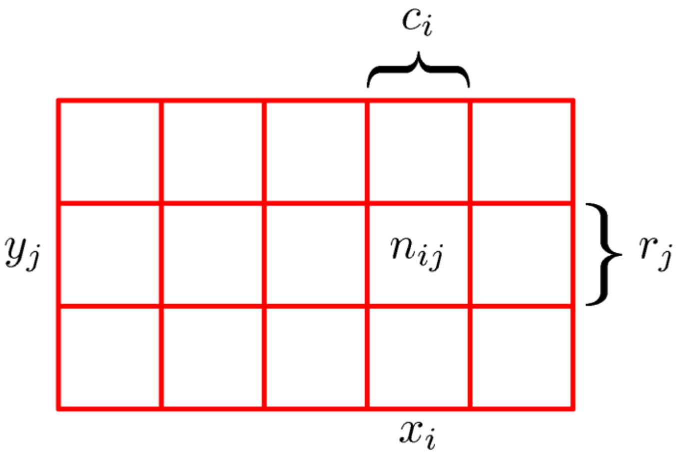
偽陽性
False positives
3%



ガンではない　ガンである

偽陰性
False negatives
10%

確率の基本法則 The rules of probability
加法定理と乗法定理 The sum and product rules



周辺確率 Marginal probability

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

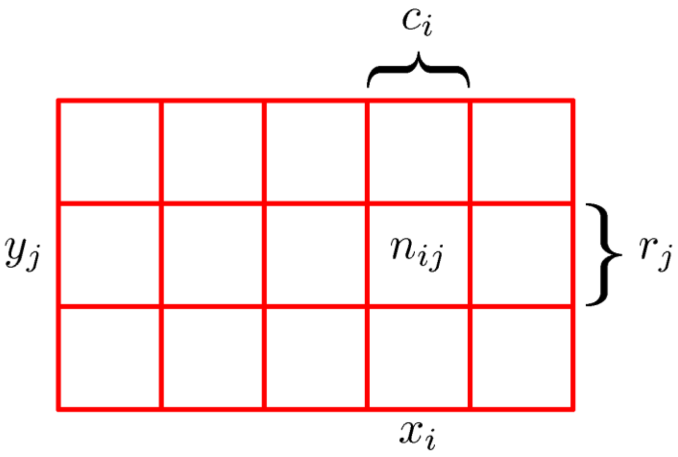
同時確率 Joint probability

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

条件付き確率 Conditional probability

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

確率の基本法則 The rules of probability
加法定理と乗法定理 The sum and product rules



加法定理 Sum rule

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

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

乗法定理 Product rule

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

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

確率の基本法則 The rules of probability
加法定理と乗法定理 The sum and product rules

加法定理 Sum rule

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

乗法定理 Product rule

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

確率の基本法則 The rules of probability
ベイズの定理 Bayes' theorem

ベイズの定理 Bayes' theorem

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

$\Leftarrow$

乗法定理 Product rule

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

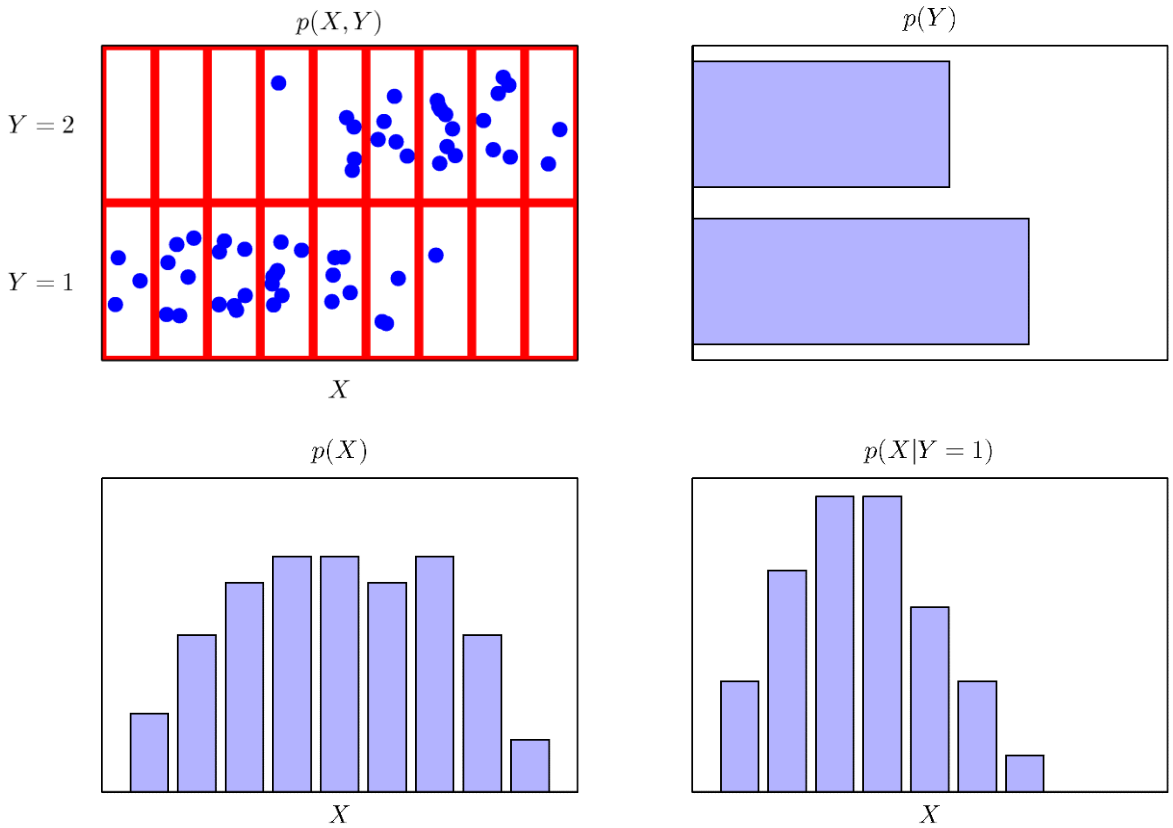
加法定理 Sum rule

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

$\Leftarrow$

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

確率の基本法則 The rules of probability
ベイズの定理 Bayes' theorem

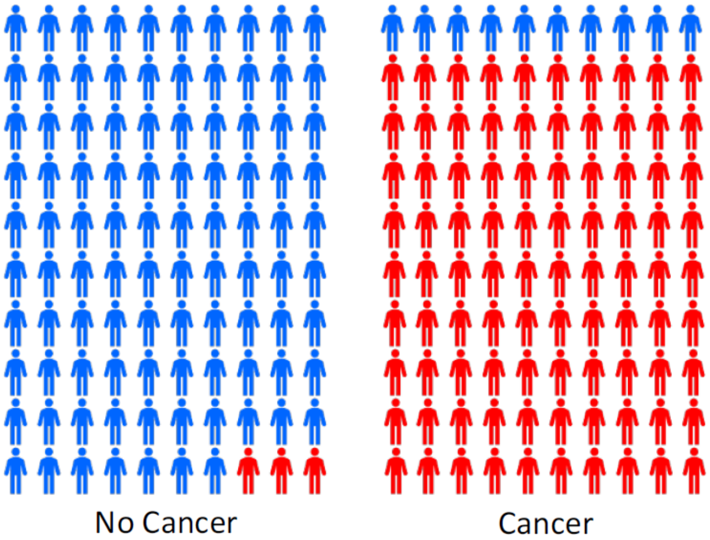


確率の基本法則 The rules of probability
医療スクリーニング再訪 Medical screening revisited

$$p(C=1) = 1/100$$ ガンである Cancer

$$p(C=0) = 99/100$$ ガンではない No cancer

$$p(T=1|C=0) = 3/100$$
$$p(T=0|C=0) = 97/100$$

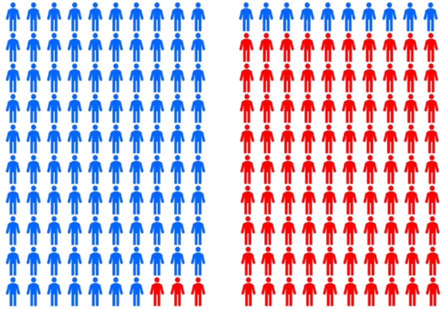


$$p(T=1|C=1) = 90/100$$
$$p(T=0|C=1) = 10/100$$

確率の基本法則 The rules of probability
医療スクリーニング再訪 Medical screening revisited

$$p(C=1) = 1/100$$ ガンである Cancer

$$p(C=0) = 99/100$$ ガンではない No cancer



$$\begin{aligned} p(T=1|C=0) &= 3/100 \\ p(T=0|C=0) &= 97/100 \end{aligned} \qquad \begin{aligned} p(T=1|C=1) &= 90/100 \\ p(T=0|C=1) &= 10/100 \end{aligned}$$

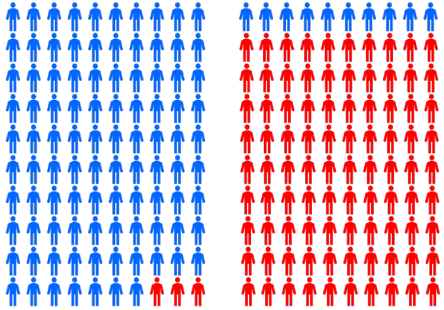
陽性のテスト結果となる確率（ランダムにテストされた被験者に対して）
Probability that someone who is tested at random will have a positive test result

$$\begin{aligned} p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1) \\ &= \frac{3}{100} \times \frac{99}{100} + \frac{90}{100} \times \frac{1}{100} = \frac{387}{10,000} = 0.0387. \end{aligned}$$

確率の基本法則 The rules of probability
医療スクリーニング再訪 Medical screening revisited

$$p(C=1) = 1/100$$ ガンである Cancer

$$p(C=0) = 99/100$$ ガンではない No cancer



$$p(T=1|C=0) = 3/100 \qquad p(T=1|C=1) = 90/100$$

$$p(T=0|C=0) = 97/100 \qquad p(T=0|C=1) = 10/100$$

テストが陽性であった場合にガンである確率
Probability that the person has cancer if a test is positive

$$\begin{aligned} p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \\ &= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} = \frac{90}{387} \simeq 0.23 \end{aligned}$$

確率の基本法則 The rules of probability
事前確率と事後確率 Prior and posterior probabilities

事前確率 Prior probability

$$p(C=1) \quad = \quad 1/100$$

事後確率 Posterior probability

$$\begin{aligned} p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \\ &= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} = \frac{90}{387} \simeq 0.23 \end{aligned}$$

確率の基本法則 The rules of probability
独立変数 Independent variables

X と Yが独立である場合：
If X and Y are independent:

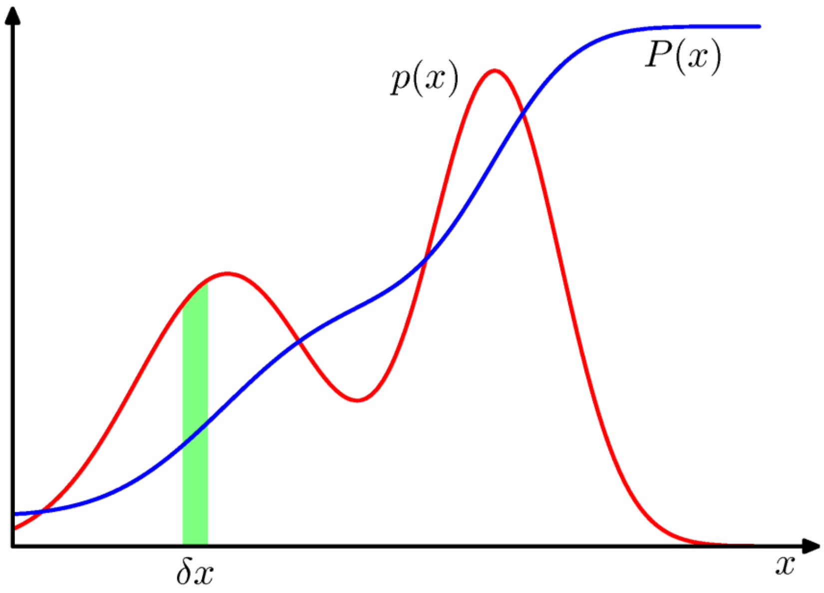
$$p(X, Y) = p(X)p(Y)$$

乗法定理より,
From the product rule,

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

確率密度 Probability densities $p(x)$

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



$P'(x) = p(x)$

x が区間 (a,b) に含まれる確率
Probability of that x will lie in an interval (a,b)

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

累積分布関数
Cumulative distribution function

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

確率密度 Probability densities $p(\mathbf{x}) = p(x_1, \ldots, x_D)$

条件 Conditions

$$\begin{aligned} p(\mathbf{x}) &\geqslant 0 \\ \int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} &= 1 \end{aligned}$$

加法定理 Sum rule

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y}$$

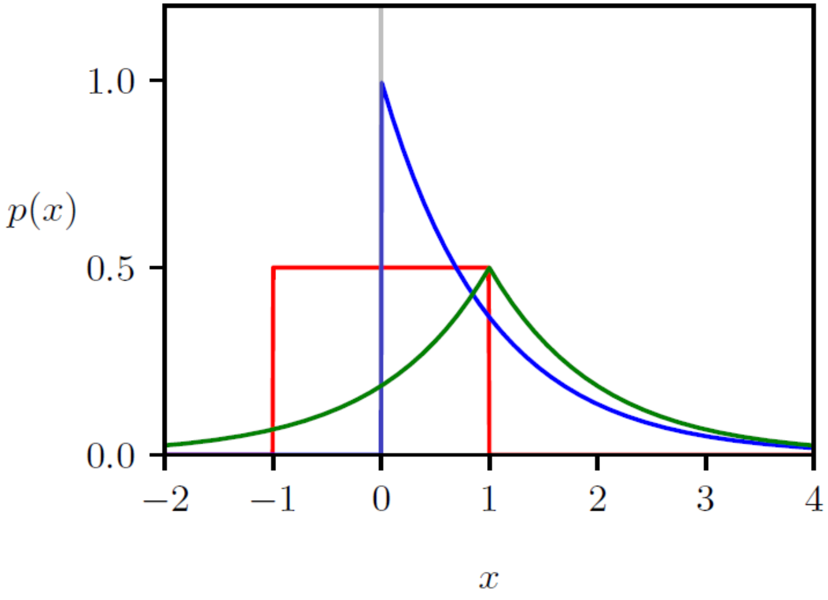
乗法定理 Product rule

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

ベイズの定理
Bayes' theorem

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y}) p(\mathbf{y})}{p(\mathbf{x})} \qquad p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}) \, \mathrm{d}\mathbf{y}$$

確率密度 Probability densities
分布の例 Example distributions



一様分布 Uniform distribution

$$p(x) = 1/(d-c), \qquad x \in (c, d).$$

指数分布 Exponential distribution

$$p(x|\lambda) = \lambda \exp(-\lambda x), \qquad x \geqslant 0.$$

ラプラス分布 Laplace distribution

$$p(x|\mu,\gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

確率密度 Probability densities
分布の例 Example distributions

ディラックのデルタ関数 Dirac delta function

$$p(x|\mu) = \delta(x - \mu)$$

経験分布 Empirical distribution

$$p(x|\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n)$$

確率密度 Probability densities
期待値と共分散 Expectations and covariances

期待値 Expectation

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

期待値の近似 Approximate expectation

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

条件付き期待値 Conditional expectation

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

確率密度 Probability densities
期待値と共分散 Expectations and covariances

分散 Variance

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

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

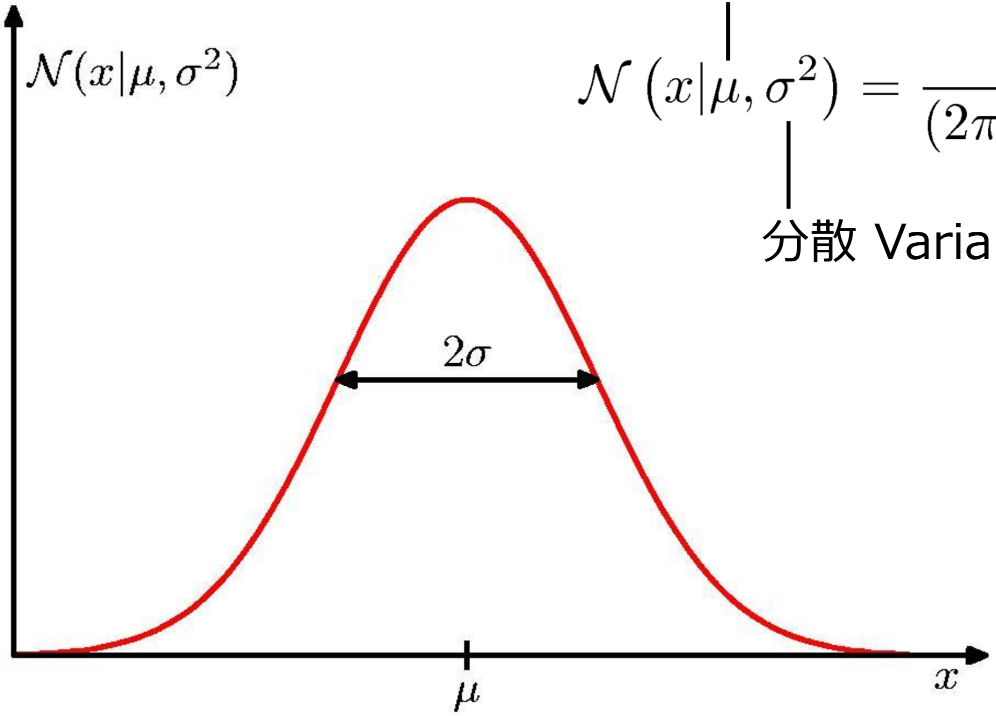
共分散 Covariance

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

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

$$\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}, \mathbf{x}]$$

ガウス分布 The Gaussian distribution



平均 Mean

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

分散 Variance

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

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

ガウス分布 The Gaussian distribution
平均と分散 Mean and variance

平均 Mean

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

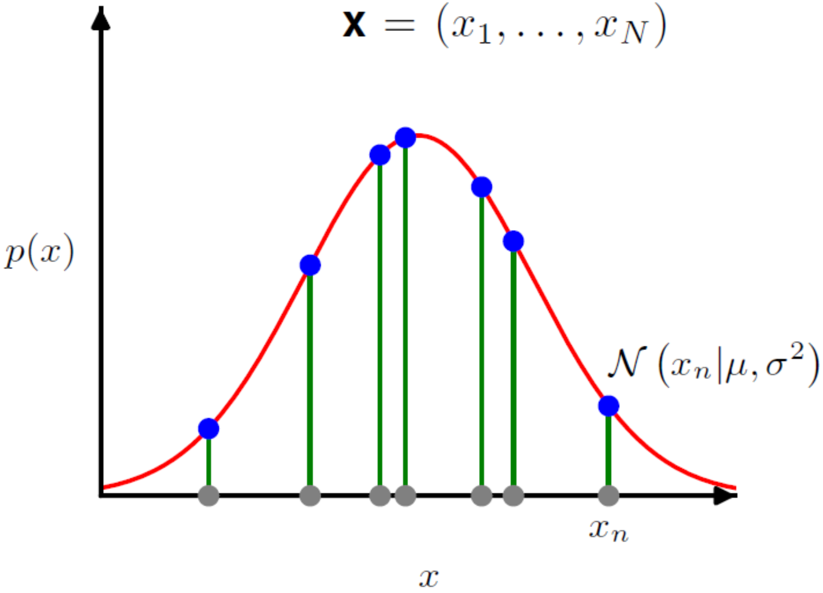
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

分散 Variance

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

ガウス分布 The Gaussian distribution
尤度関数 Likelihood function

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



ガウス分布 The Gaussian distribution
尤度関数 Likelihood function

対数尤度の最大化 Maximization of log likelihood

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

最尤推定解 Maximum likelihood solution

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

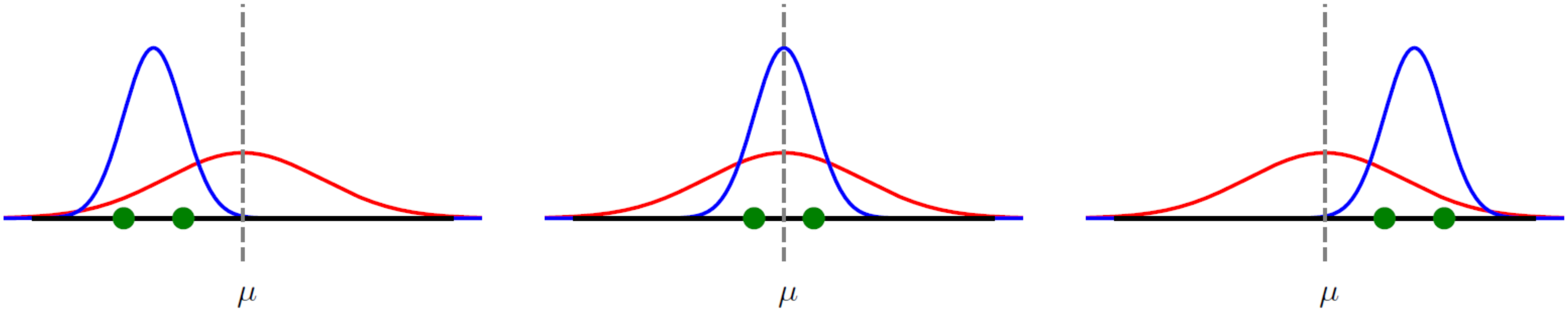
ガウス分布 The Gaussian distribution
最尤推定のバイアス Bias of maximum likelihood

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

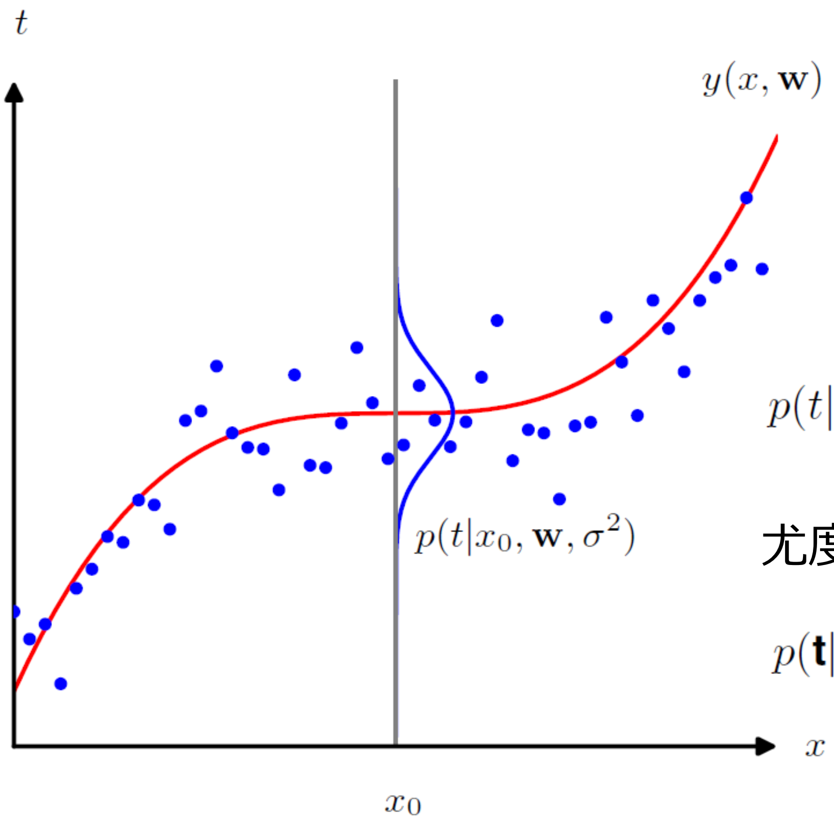
$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

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

$$\mathbb{E}\left[\widetilde{\sigma}^2\right] = \sigma^2$$



ガウス分布 The Gaussian distribution
線形回帰 Linear regression



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

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

$$p(t|x, \mathbf{w}, \sigma^2) = \mathcal{N}\left(t|y(x, \mathbf{w}), \sigma^2\right)$$

尤度関数 Likelihood function

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

ガウス分布 The Gaussian distribution
線形回帰 Linear regression

尤度関数 Likelihood function

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

二乗和誤差 Sum-of-squares error

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

対数尤度 Log likelihood

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

$\mathbf{w}_{\mathrm{ML}}$ は $E(\mathbf{w})$ を最小化することで得られる.

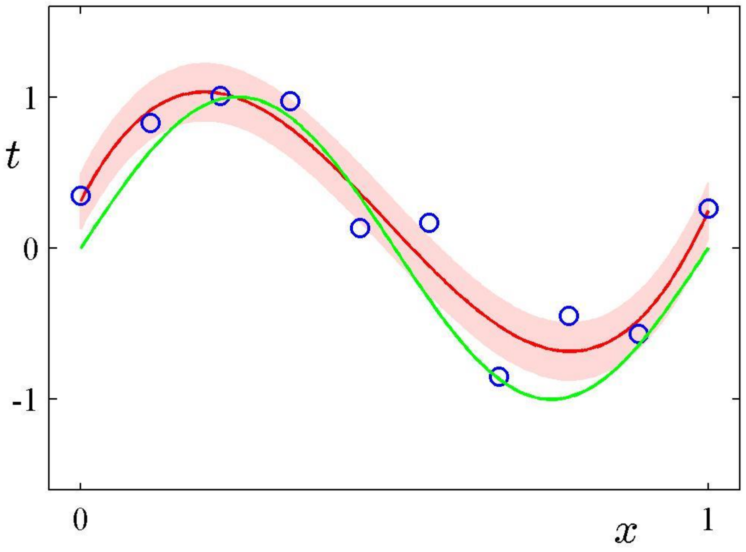
$\mathbf{w}_{\mathrm{ML}}$ is obtained by minimizing $E(\mathbf{w})$.

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

ガウス分布 The Gaussian distribution
線形回帰 Linear regression

予測分布 Predictive distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \sigma^2_{\mathrm{ML}}\right)$$



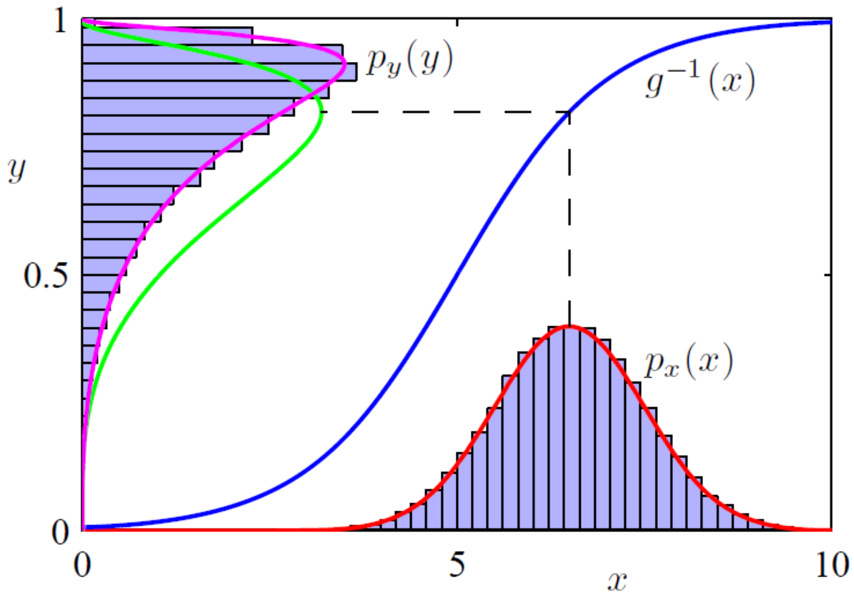
密度の変換 Transformation of densities

変数変換 Change of variables

$$x = g(y)$$

$$p_x(x)\delta x \simeq p_y(y)\delta y$$

$$\Rightarrow \quad \begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\left|\frac{\mathrm{d}g}{\mathrm{d}y}\right| \end{aligned}$$



$$x = g(y) = \ln(y) - \ln(1-y) + 5$$

$$y = g^{-1}(x) = \frac{1}{1+\exp(-x+5)}$$

密度の変換 Transformation of densities
多変量分布 Multivariate distributions

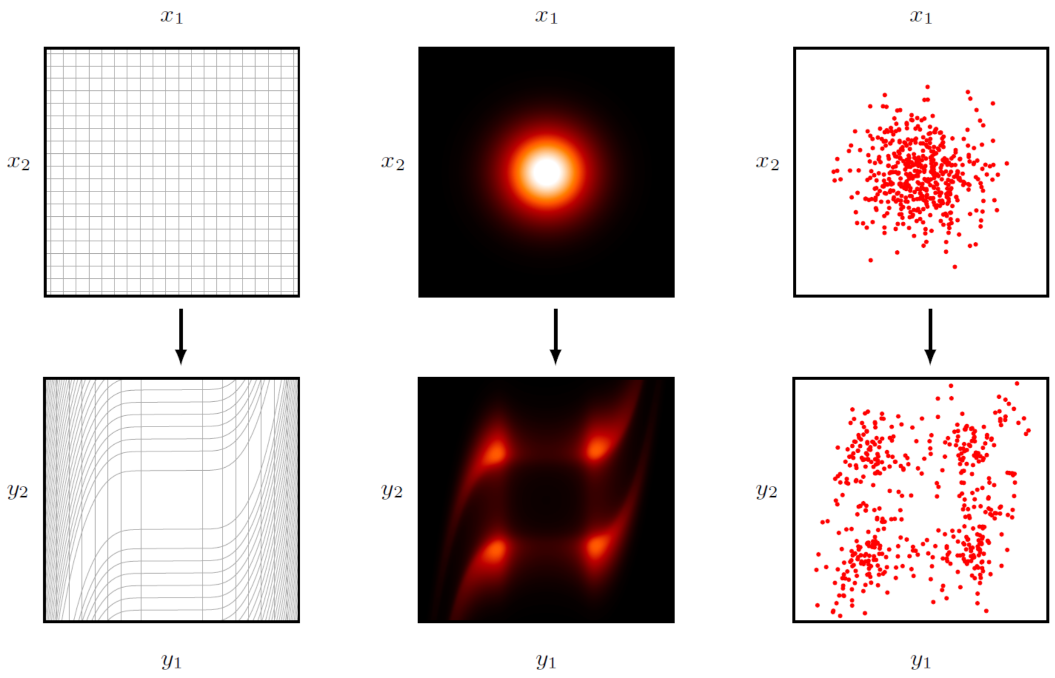
変数変換 Change of variables

$$\mathbf{x} = \mathbf{g}(\mathbf{y})$$

$$\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \quad \mathbf{y} = (y_1, \ldots, y_D)^{\mathrm{T}}$$

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}) \left|\det \mathbf{J}\right|$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D}{\partial y_1} & \cdots & \frac{\partial g_D}{\partial y_D} \end{bmatrix}$$



$$\begin{aligned} y_1 &= x_1 + \tanh(5x_1) \\ y_2 &= x_2 + \tanh(5x_2) + \frac{x_1^3}{3} \end{aligned}$$

情報理論 Information theory
エントロピー Entropy

自己情報量 Self-information $$h(x) = -\log_2 p(x)$$

エントロピー Entropy $$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

同じ確率で8つの状態をもつ確率変数 x に対しては,

For a random variable x having 8 possible states, each of which is equally likely,

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

情報理論 Information theory
エントロピー Entropy

自己情報量 Self-information $h(x) = -\log_2 p(x)$

エントロピー Entropy $\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$

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

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

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

$$= 2 \text{ bits}$$

情報理論 Information theory
物理学の観点 Physics perspective

N 個の物体の，箱の入れ方への総数
The total number of ways of allocating the N objects to the bins

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

i 番目の箱には n_i 個の物体
n_i objects in the i th bin

エントロピー Entropy

$$\ln N! \simeq N \ln N - N \qquad \sum_i n_i = N$$

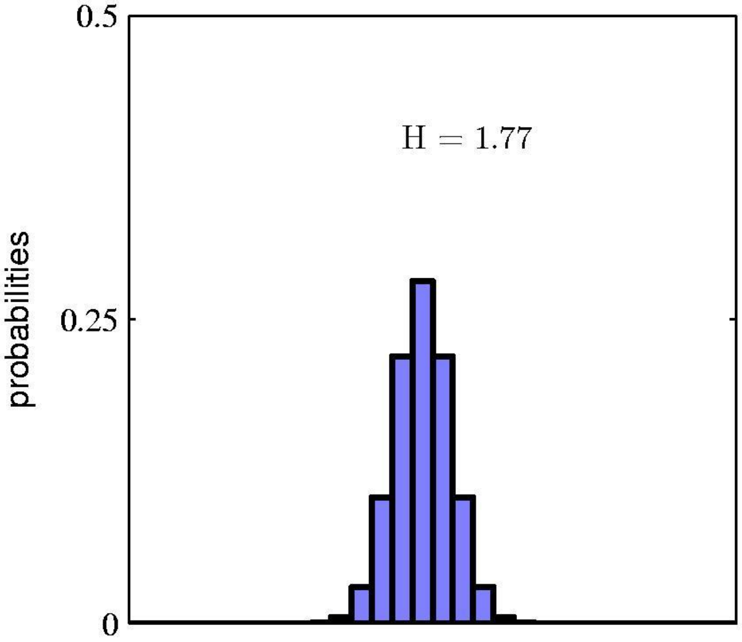
$$\mathrm{H} = \frac{1}{N} \ln W \simeq -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

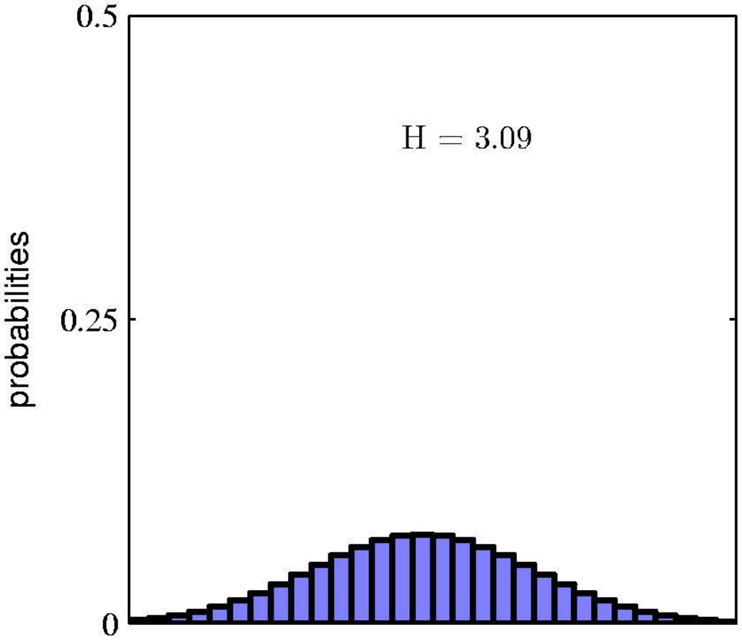
エントロピーは次の場合のとき最大
Entropy is maximized when

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

M: 箱の総数
Total number of bins

情報理論 Information theory
物理学的な観点 Physics perspective





情報理論 Information theory
微分エントロピー Differential entropy

微分エントロピー Differential entropy

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

x は連続変数
x is a continuous variable

情報理論 Information theory
エントロピー最大化 Maximum entropy

微分エントロピーは以下のとき最大
Differential entropy is maximized when

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

このとき微分エントロピーの値は
where the differential entropy is

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

情報理論 Information theory
カルバック–ライブラー距離　Kullback–Leibler divergence

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

相対エントロピー
Relative entropy

性質 Properties　$\mathrm{KL}(p\|q) \geqslant 0$　$\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$

KL距離の最小化は尤度の最大化と等価.
Minimizing the KL divergence is equivalent to maximizing the log likelihood.

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

情報理論 Information theory
条件付きエントロピー Conditional entropy

条件付きエントロピー Conditional entropy

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

条件付きエントロピーは以下の関係を満たす：
The conditional entropy satisfies the relation:

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

情報理論 Information theory
相互情報量 Mutual information

相互情報量 Mutual information

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

相互情報量は以下の関係を満たす：
The mutual information satisfies the relation:

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

ベイズ確率 Bayesian probabilities

コイン投げの結果 Results of coin flipping

60%

40%

事後確率
Posterior probability

事前確率 Prior probability

50%

50%

ベイズ確率 Bayesian probabilities
モデルパラメータ Model parameters

ベイズの定理より From Bayes' theorem

尤度関数 Likelihood function

事後分布 Posterior distribution

事前分布 Prior distribution

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

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

頻度論の立場 Frequentist perspective

vs.

ベイズの立場 Bayesian perspective

ベイズ確率 Bayesian probabilities
正則化 Regularization

最大事後確率推定 Maximum a posteriori (MAP) estimation

以下の式を最小化 Minimizing the following equation

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) - \ln p(\mathbf{w}) + \ln p(\mathcal{D}) \quad \Leftarrow \quad p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$\uparrow$$

$$p(\mathbf{w}|s) = \prod_{i=0}^{M} \mathcal{N}(w_i|0, s^2) = \prod_{i=0}^{M} \left(\frac{1}{2\pi s^2}\right)^{1/2} \exp\left\{-\frac{w_i^2}{2s^2}\right\}$$

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) + \underbrace{\frac{1}{2s^2}\sum_{i=0}^{M} w_i^2}_{\text{正則化項 Regularization term}} + \text{const}$$

ベイズ確率 Bayesian probabilities
正則化 Regularization

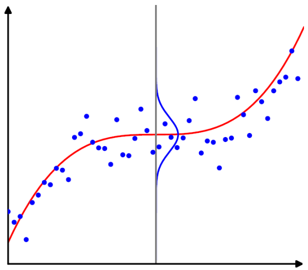
最大事後確率推定 Maximum a posteriori (MAP) estimation

以下の式を最小化 Minimizing the following equation

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) + \frac{1}{2s^2}\sum_{i=0}^{M} w_i^2 + \text{const}$$

対数尤度 Log likelihood

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



誤差関数
Error function

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

ベイズ確率 Bayesian probabilities
ベイズ機械学習 Bayesian machine learning

予測分布 Predictive distribution

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

$\mathbf{w}$ を $p(\mathbf{w}|\mathcal{D})$ を用いて積分消去（周辺化）

$\mathbf{w}$ is integrated out using $p(\mathbf{w}|\mathcal{D})$ (marginalization)