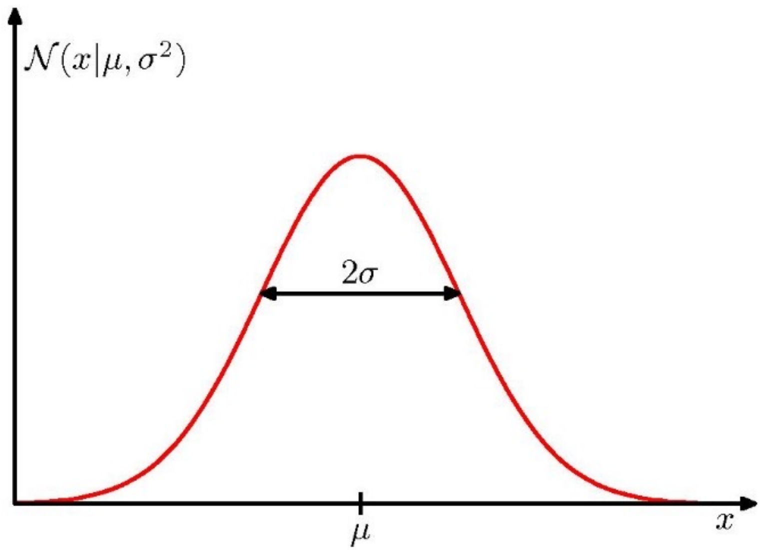
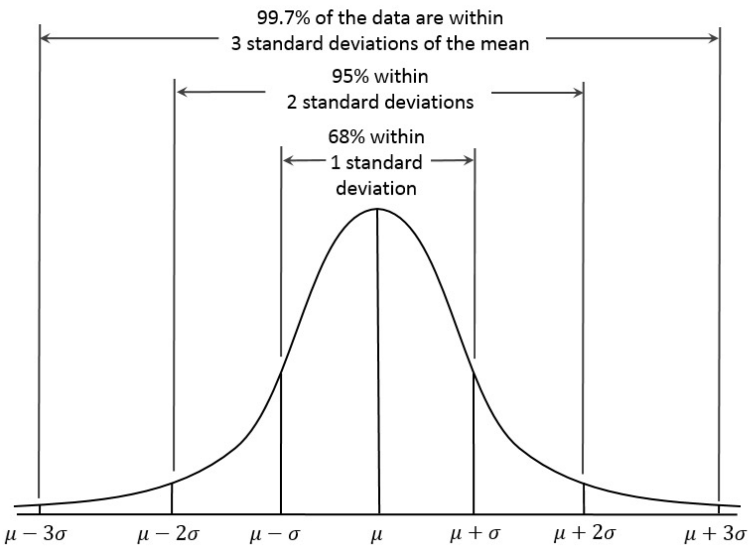
The Gaussian Distribution

- The Gaussian distribution
  - $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \times \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

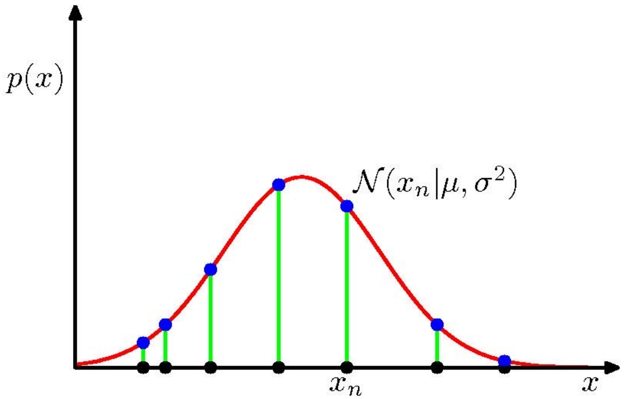


- Governed by two parameters:
  - μ is the mean
  - σ^2 is the variance
- Two other parameters:
  - σ is called the standard deviation
  - $\beta = 1/\sigma^2$ is called the precision
- Plots of Gaussian distribution



The Gaussian Distribution

- Facts about Gaussian distribution:
- Nonnegative: $\mathcal{N}(x|\mu,\sigma^2) > 0.$
- Sum to 1: $\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1.$
- Expectation: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu.$
- Variance: $\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$
- Likelihood of for data points x_n:



The Gaussian Distribution

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

- The D $\times$D matrix Σ is the covariance

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

- cov[**x**] = cov[**x**, **x**]

- Assume that observations **x** are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown.

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

MLE

- For the moment, assume that the mean μ and variance σ^2 are unknown **constants**. Can we learn the mean and the variance from the observations?

- We can learn the mean and the variance by maximizing the likelihood function

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

- This is called *maximum likelihood estimation* (**MLE**)

- However, the optimization problem is complicated…

MLE

- Recall that the logarithm function is a monotonically increasing functions. Hence, it suffices to maximize the natural log of the likelihood, over the same set of variables.

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

- This is a function in two variables. Is it a concave function?

- For each fixed σ^2, we can maximize the function over μ, resulting

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

- Note that this solution is independent of σ^2.

MLE

- If we plug in μ_{ML} in the log of the likelihood, we obtain a function in σ^2 .

- If we take the first order derivative and set it to zero, we get

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

- Does the pair (μ_{ML}, σ^2_{ML}) guarantee maximum of the likelihood?

- The second order derivate with respect to σ^2 is <0 at σ^2_{ML}, provided that σ^2_{ML} is positive.

- This is guaranteed when not all x_n are equal. When all x_n are equal, the training set does not make sense.

MLE

- When σ^2 goes to ∞, the function goes to $-\infty$. Hence the maximum is achieved at some (finite valued) point. At that point, the first order derivate with respect to σ^2 must be equal to 0.

- However, σ^2_{ML} is the unique value for the second order derivate to become 0.

- Therefore the pair (μ_{ML}, σ^2_{ML}) guarantees maximum of the likelihood

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

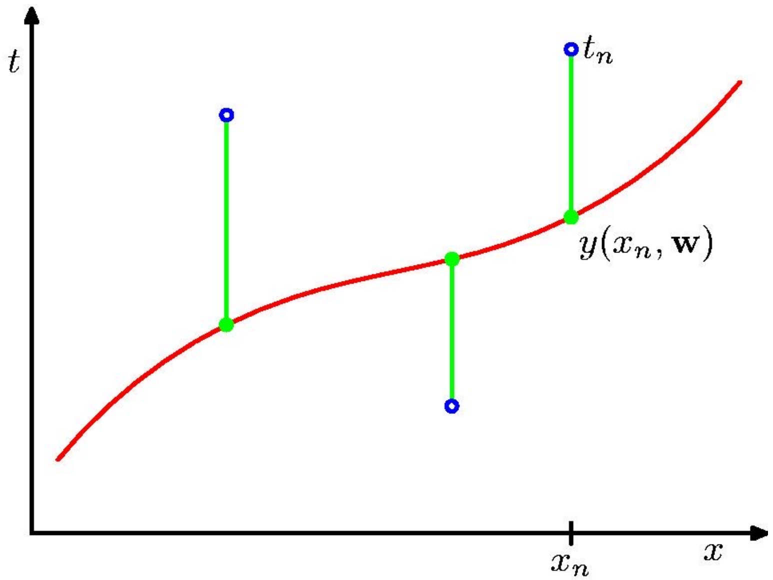
MLE

- MLE: $$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$
- We can view the estimations themselves as random variables.
- Furthermore, we have

$$\begin{aligned}
\mathbb{E}[\mu_{\mathrm{ML}}] &= \mu \\
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned}$$

Curve Fitting Re-visited

- Sum-of-squares error function: $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$

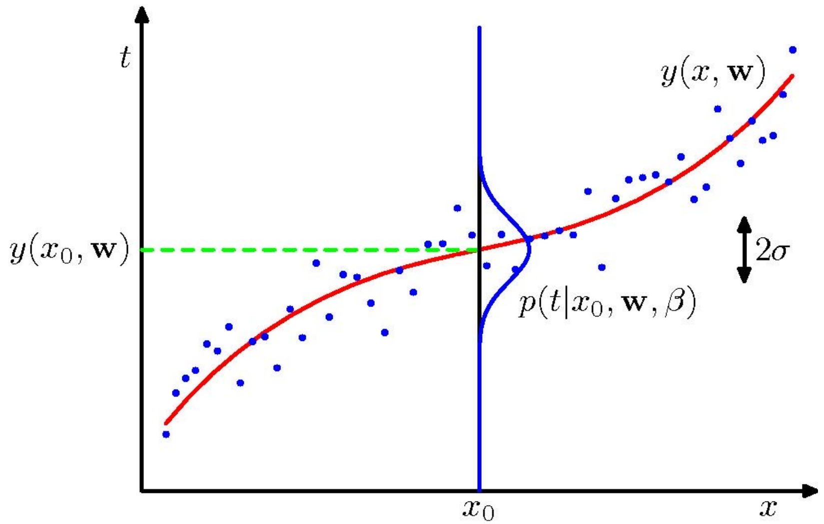


- Minimize E(w) to determine the optimal parameters w.

Curve Fitting Re-visited

- Assume that given the value of x, the corresponding value of t has a Gaussian distribution, with a mean equal to the value of y(x, w), and a precision β.
- Thus $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$
- Likelihood:

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



Curve Fitting Re-visited

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

- MLE solution for the mean is equivalent to minimizing the sum-of-squares

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

- In addition, MLE also provides an estimation of the precision

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

Curve Fitting Re-visited

- Using the training data set, MLE computes w_{ML} and β^{-1}_{ML}

- For any new value of x, the distribution of t is given by

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

- This is a *predictive distribution*.

MAP

- Assume a prior distribution over the coefficients w.

- Consider a Gaussian distribution with mean equal to 0

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

- Recall (p. 5), and D=M+1,

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

- The posterior distribution for w is

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

MAP

- *Maximum posterior* (MAP):

- ln {p(t|x, w, β) p(w|α)} = ln p(t|x, w, β) + ln p(w|α), where

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

- Hence MAP minimizes

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

- MAP is equivalent to minimizing the regularized sum-of-squares, with $\lambda=\alpha/\beta$.