CSE 575
Statistical Machine Learning

Lecture 4

YooJung Choi

Fall 2022

Monty Hall - variation

- After you choose a door, the host reveals one of the others *at random.* If it happens to reveal a goat, should you stick with your choice or switch?
- W.l.o.g. say you chose door 1, and the host revealed door 2.
- $p(C = 3|H_1 = 2)$?

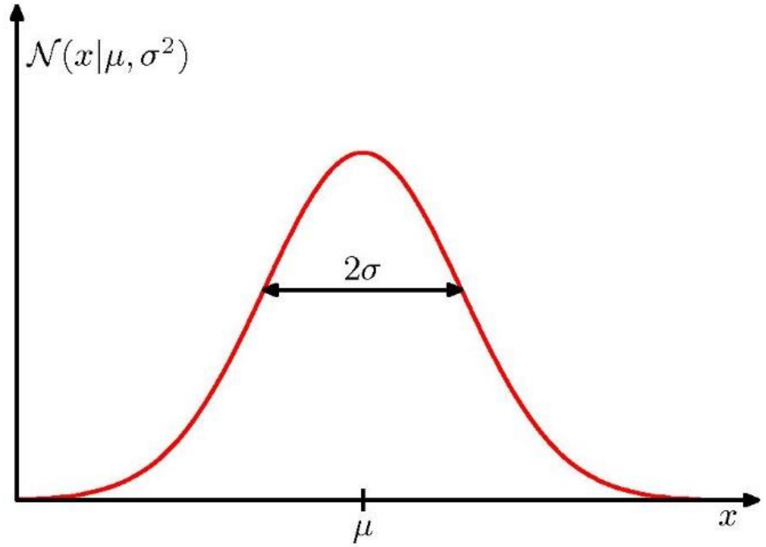
$$\frac{p(H_1 = 2|C = 3) \cdot p(C = 3)}{\sum_i p(H_1 = 2|C = i) \cdot p(C = i)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \frac{1}{2}$$

The Gaussian distribution

- Normal / Gaussian distribution:

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

  parameterized by mean μ, variance σ^2



- Multivariate Gaussian over a D-dimensional vector **x**:

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

Likelihood

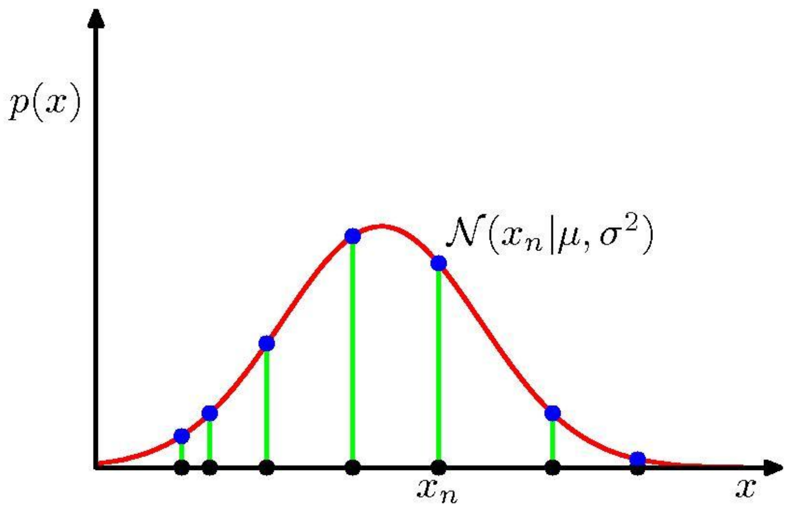
- Probability density of a single point *x*:

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

- Likelihood given a dataset of observations $\mathbf{x} = (x_1, \dots, x_N)^T$

  (assume i.i.d)

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



Maximum likelihood estimation (MLE)

- 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

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

over μ, σ^2

Maximum likelihood estimation (MLE)

- Because logarithm is a monotonically increasing function, we can equivalently maximize the *log-likelihood*

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

- Benefit 1: simplified mathematical analysis
- Benefit 2: numerical stability

Maximum likelihood estimation (MLE)

- For a fixed σ^2, we can maximize w.r.t. μ to get:

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

Note:
independent of σ^2

- Similarly, we can fix $\mu = \mu_{ML}$ and maximize w.r.t. σ^2:

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

- Note how the joint optimization w.r.t. μ and σ^2 can be decoupled

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

- Fix σ^2. Optimize wrt μ to get the sample mean:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \quad \Rightarrow \quad \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} s_n$$

- Note: $\frac{\partial^2 l}{\partial \mu^2} < 0$ and the log likelihood is concave on μ

- Replace $\mu = \mu_{ML}$. Optimize wrt σ^2 to get the sample variance:

$$\frac{\partial l}{\partial \sigma} = \frac{1}{\sigma^3}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 - \frac{N}{\sigma} = 0 \quad \Rightarrow \quad \sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

- (assuming σ is non-zero)

Maximum likelihood estimation (MLE)

- *Soundness check: Does the pair* $(\mu_{ML}, \sigma^2_{ML})$ *guarantee maximum likelihood?*
- When σ^2 goes to ∞, the likelihood goes to -∞. 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.
- But σ^2_{ML} is the unique value for the derivate to become 0.
- The second-order derivative w.r.t. σ^2 is < 0 at σ^2_{ML} as long as not all x_n are equal.
- Thus $(\mu_{ML}, \sigma^2_{ML})$ guarantees maximum of the likelihood

Maximum likelihood estimation (MLE)

- MLE solutions are functions of the dataset values

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

- i.e. they are also random variables
- Furthermore, we have

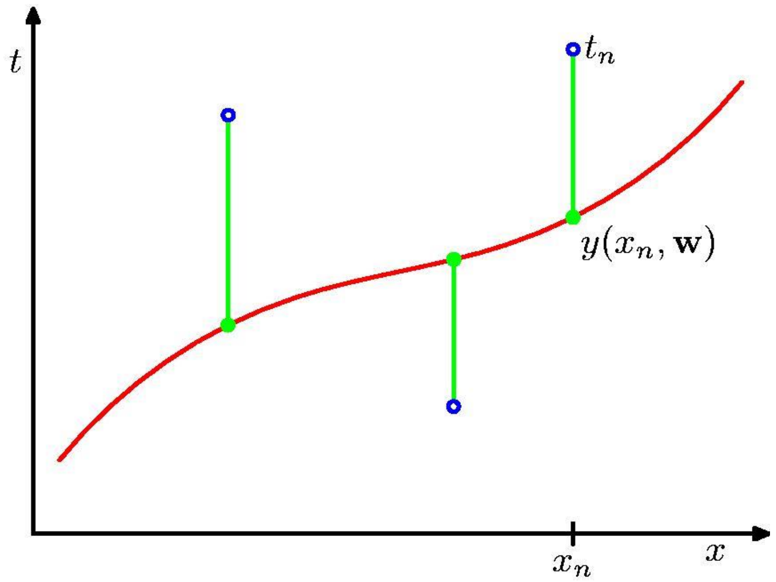
$$\begin{aligned}\mathbb{E}[\mu_{\mathrm{ML}}] &= \mu \\ \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \left(\frac{N-1}{N}\right)\sigma^2\end{aligned}$$

Curve fitting re-visited

- Recap: Given a training data with *N* input values $\mathbf{x} = (x_1, \dots, x_N)^T$ and corresponding target values $\mathbf{t} = (t_1, \dots, t_N)^T$, fit a polynomial $y(\mathbf{x}, \mathbf{w})$
- Sum-of-squares error function:

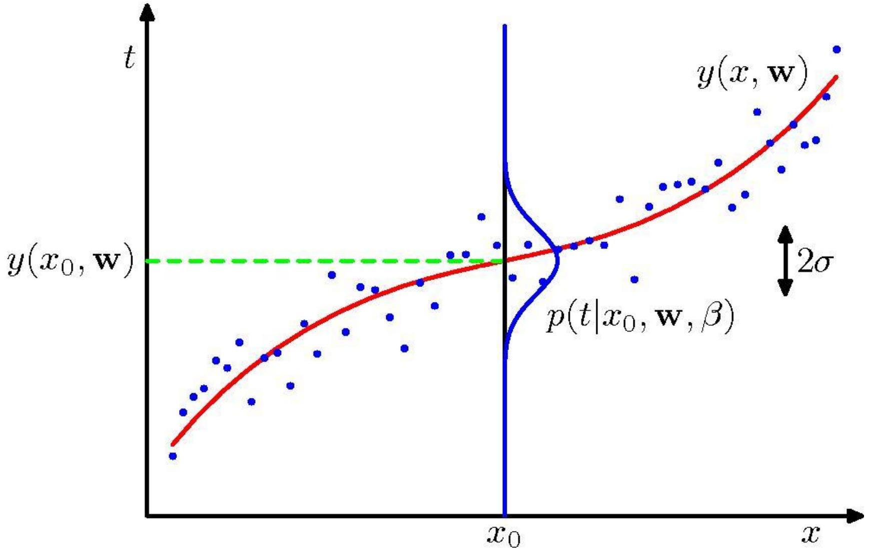
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

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



Curve fitting: probabilistic view

- Assume that given the value of x, the corresponding value of t has a Gaussian distribution with mean $y(\mathbf{x}, \mathbf{w})$
- Thus, $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$
- $\beta = 1/\sigma^2$ is called a precision parameter



Curve fitting: probabilistic view

- Likelihood:

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

- Max likelihood is equivalent to min negative log-likelihood:

$$\begin{aligned} &\operatorname{argmax}_{\mathbf{w},\beta}\, p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \\ &= \operatorname{argmax}_{\mathbf{w},\beta}\, \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \\ &= \operatorname{argmin}_{\mathbf{w},\beta}\, -\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \\ &= \operatorname{argmin}_{\mathbf{w},\beta}\, \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2} \ln \beta + \frac{N}{2} \ln(2\pi) \end{aligned}$$

Curve fitting: probabilistic view

$$\text{argmin}_{\mathbf{w},\beta} \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2} \ln \beta + \frac{N}{2} \ln(2\pi)$$

- Optimizing w.r.t. polynomial coefficients **w:** equivalent to minimizing the sum-of-squares

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

- MLE solution also provides an estimation of 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: probabilistic view

- After determining $\boldsymbol{w}_{ML}, \beta_{ML}$ from data, how do we make predictions on new x?
- For each new x, we now have a distribution over t:

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

- *“Predictive distribution”*

Curve fitting: introducing prior

- Recall from Bayes' theorem: $Posterior \propto Likelihood \times Prior$
- In general, for parameters $\boldsymbol{\theta}$ and data $\mathcal{D}$:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

- Let's assume a prior distribution over the polynomial coefficients $\boldsymbol{w}$
- E.g. a Gaussian with mean = 0

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

Curve fitting: introducing prior

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

- Recall the multivariate Gaussian distribution: (D=M+1)

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

Note: diagonal covariance matrix
=> independent parameters

- 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

- We can now choose the most probable value of **w** given data
- *Maximum Posterior (MAP)*: maximize the posterior distribution (minimize its negative log)

  $$\min -\ln(p(t|x, \boldsymbol{w}, \beta) \times p(\boldsymbol{w}|\alpha)) = \min(-\ln p(t|x, \boldsymbol{w}, \beta) - \ln p(\boldsymbol{w}|\alpha))$$

- Equivalent to minimizing

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

- *MAP is equivalent to minimizing the L2-regularized sum-of-squares, with $\lambda=\alpha/\beta$*