Gaussian Processes and Bayesian Methods

CAC 고등과학원 여름학교
2022. 6. 27 - 7. 1. (Mon.-Fri.)

Yung-Kyun Noh (노영균)
Hanyang University &
Korea Institute for Advanced Study


classifier
f(x;θlearned)
labe
DATA
classification
BIG
DATA
y(x;D)=f(x;θlearned)
LAB OF ADVANCED
APPLICATION FOR
INTELLIGENCE SYSTEMS

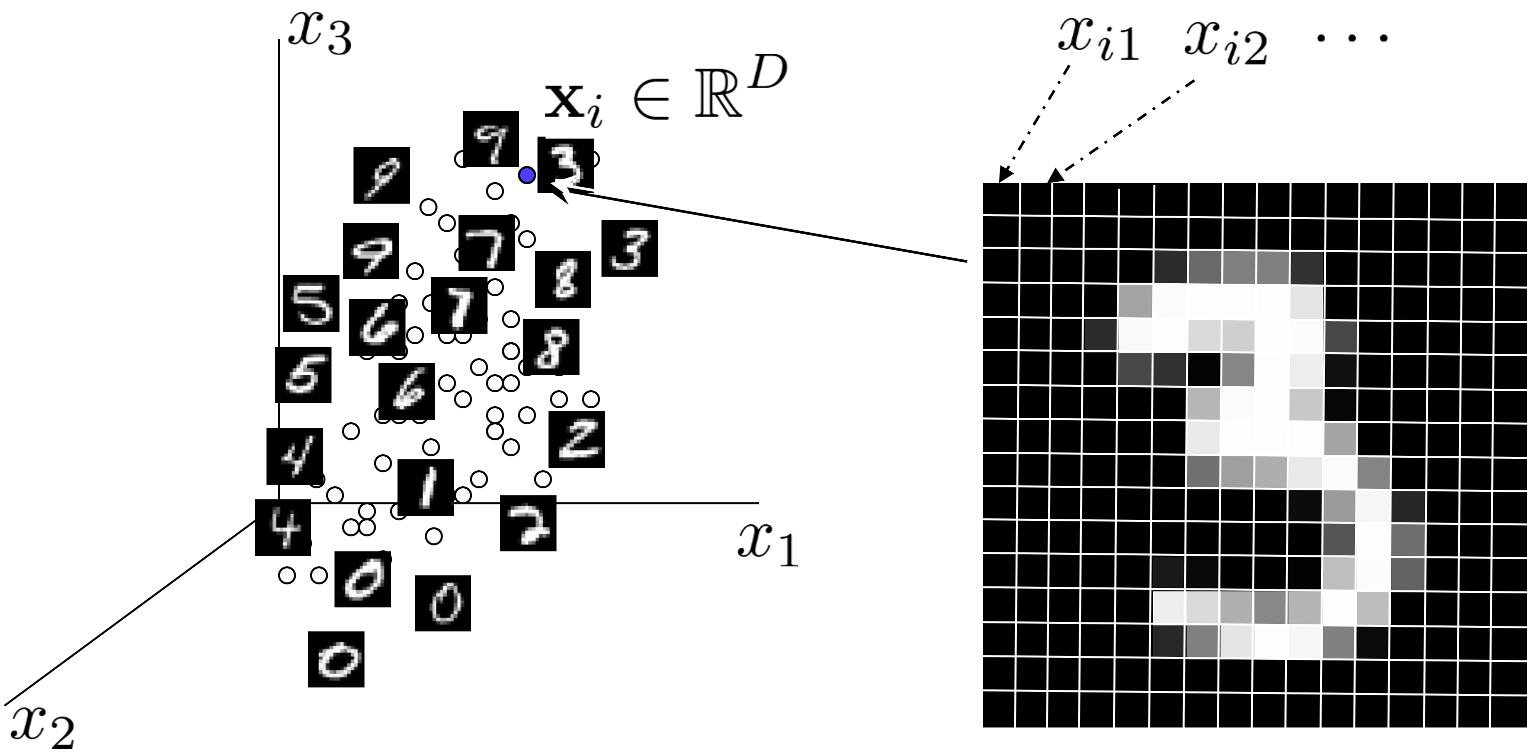
- Representation of data

- What does it mean by learning from data
  – Computer's learning method
  – Generalization

- Gaussians and parameter estimation

Representation of Data

Data Space



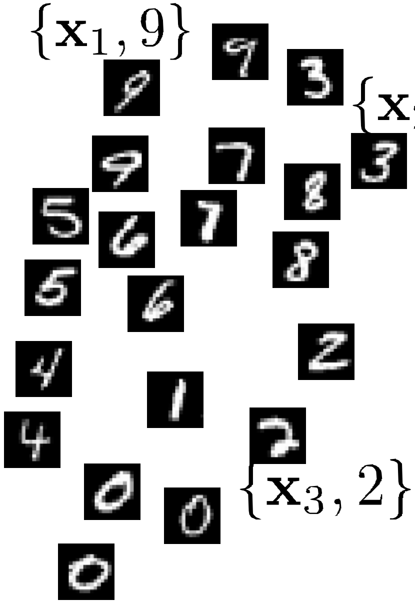
$$\mathbf{x}_i = [x_{i1}, \ldots, x_{iD}]$$
$$= [0, 0, 0, 202, 250, \ldots]$$

- Each datum is one point in a data space

Machine Learning is All About "Data" and Generalization

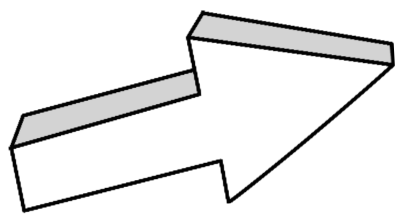
- Prediction Pipeline

Collected Data

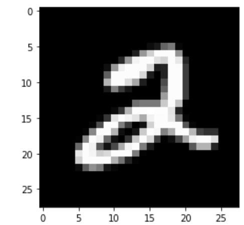


Model

$$\mathcal{H} = \{f(\mathbf{x}; \theta) | \theta \in \Theta\}$$



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



2 or

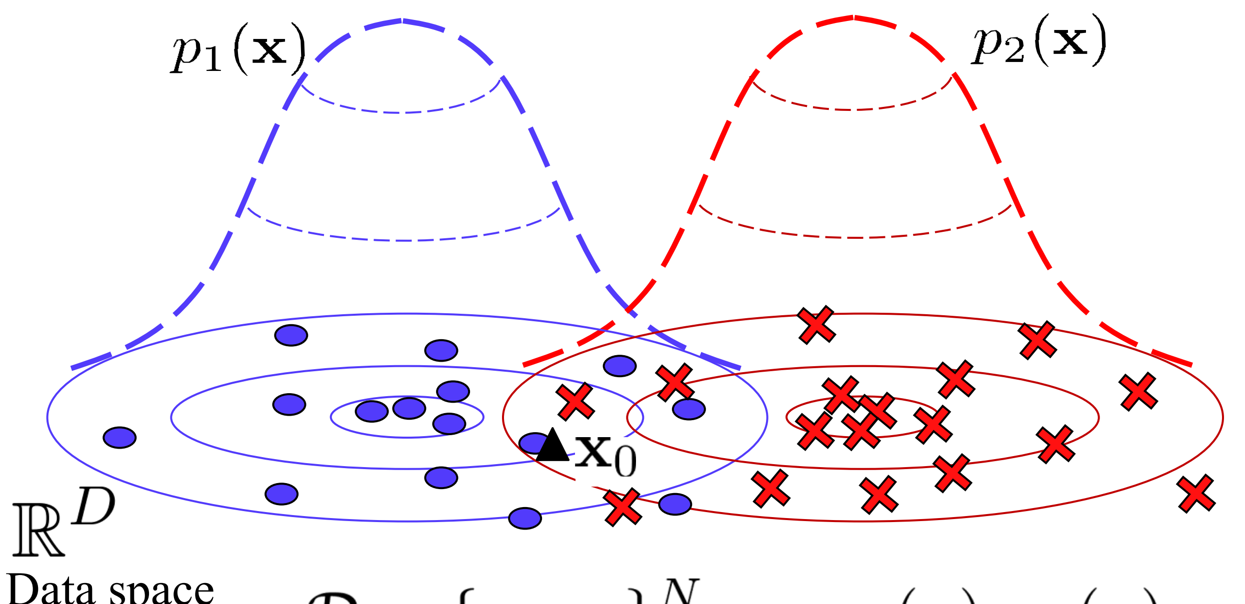
[0,0,**1**,0,0,0,0,0,0,0]

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

$$\mathbf{x}_i \in \mathbb{R}^D, y_i \in \{0, 1, \ldots, 9\}$$

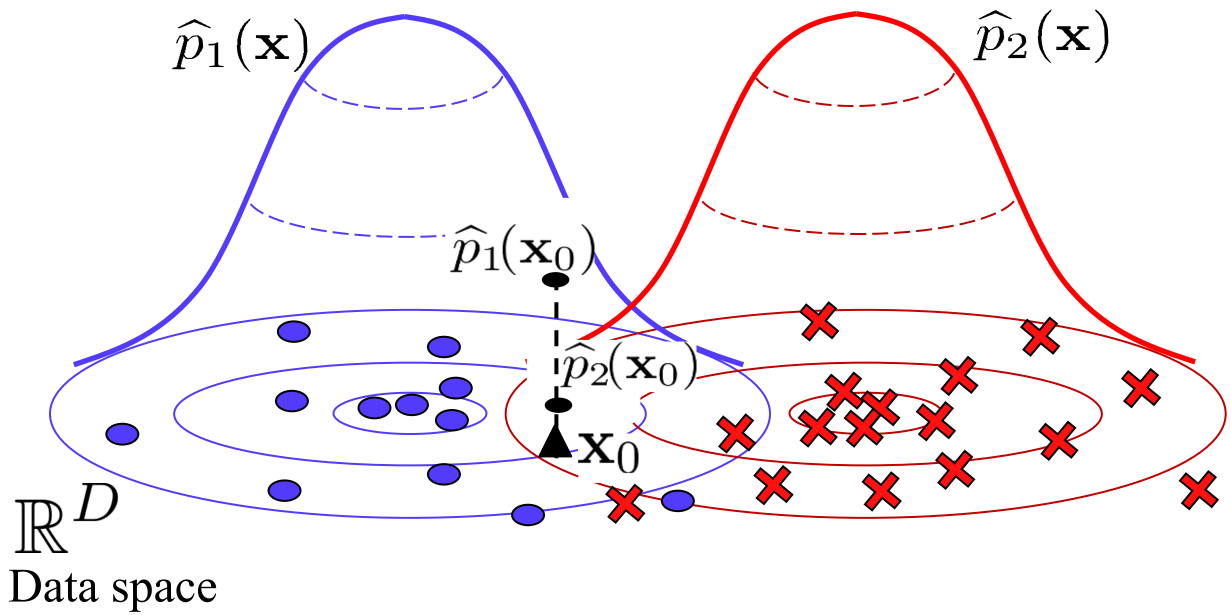
Machine Learning Assumptions

- What does make prediction possible?
  – Assumptions on the (true / data generating / underlying) probability density functions



$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$

Bayes Classification and Generalization



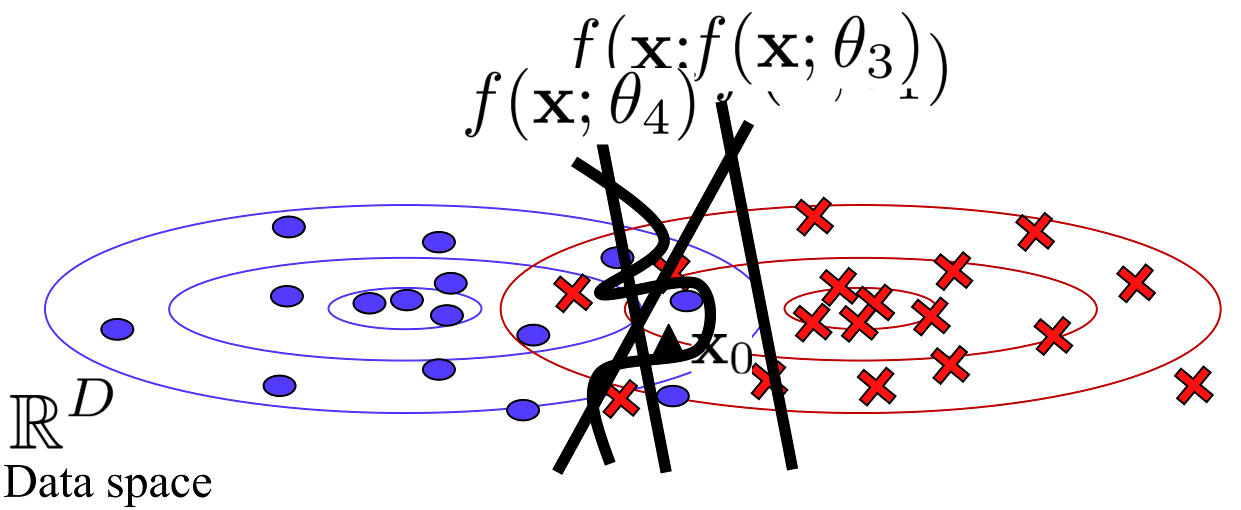
$$\widehat{p}_1(\mathbf{x}_0) \gtrless \widehat{p}_2(\mathbf{x}_0)$$

Misclassification rate $> R^*$ (Misclassification using true density functions)

Generative vs. Discriminative Models

- Discriminative Models – Models on posterior or discrimination function

$$\mathcal{H} = \left\{ f(\mathbf{x};\theta) \in \{1,2\} \;\middle|\; \theta \in \Theta \right\}$$



$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$

Models in Science vs. Models for Prediction

- Richard. P. Feynman (1998)
  - the more specific a rule is, the more interesting it is. The more definite the statement, the more interesting it is to test.
- George. E. P. Box (1979)
  - All models are wrong but some are useful
  - For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

Model for Prediction

- Set of candidate functions

$$\mathcal{H} = \{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_{N_{\mathcal{H}}}(\mathbf{x})\}$$

In general, $N_{\mathcal{H}}$ is infinite.

  - $h_i(\mathbf{x})$ can be a prediction function

$$h_i(\mathbf{x}) \rightarrow y \qquad f(\mathbf{x}) = h_i(\mathbf{x})$$

  - $h_i(\mathbf{x})$ can be a probability density

$$h_i(\mathbf{x}) \rightarrow p(\mathbf{x}) \qquad f(\mathbf{x}) = \frac{h_i(\mathbf{x})}{h_i(\mathbf{x}) + h_j(\mathbf{x})}$$

Quantify the Evaluation (Use Data)

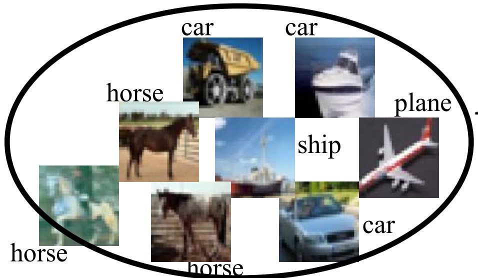
- Measure of quality: expected loss

$$L = \mathbb{E}_P[l(y, f(\mathbf{x}))] \qquad l(y, y'): \text{loss function}$$

- Estimated error

$$\hat{L} = \sum_n l(y_n, f(\mathbf{x}_n)), \qquad f(\mathbf{x}) \in \mathcal{H}$$

- Examples
  - Classification $\quad l(y, f(\mathbf{x})) = \mathbb{I}(y \neq f(\mathbf{x}))$
  - Regression $\quad l(y, f(\mathbf{x})) = ||y_n - f(\mathbf{x}_n)||^2$
  - Clustering $\quad l(f(\mathbf{x})) = \min_{c_n \in \mathcal{C}} ||c_n - f(\mathbf{x}_n)||^2$



X_{train}

$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$

$f(\mathbf{x};\theta)$

Learn θ

$f(\mathbf{x};\theta) \in \mathcal{H}$

classifier

New
unseen
data

$f(\mathbf{x};\theta_{learned})$

label
(a horse)

$$y(\mathbf{x};\mathcal{D}) = f(\mathbf{x};\theta_{learned}(=\widehat{\theta}))$$

[Based on the assumption of regularity]

Consistent Learner

- Model $\mathcal{H}$ satisfies

$$\widehat{L} \underset{N \to \infty}{\longrightarrow} L$$

$$P\{\sup_{f \in \mathcal{H}} (L(f) - \widehat{L}(f)) > \epsilon\} \to 0 \quad \text{for} \quad \epsilon > 0$$

<Uniform convergence>

- Caution:
  - The definition of consistency is *not*

$$\widehat{L}(f) \to L(f) \quad \text{for} \quad f \in \mathcal{H}$$

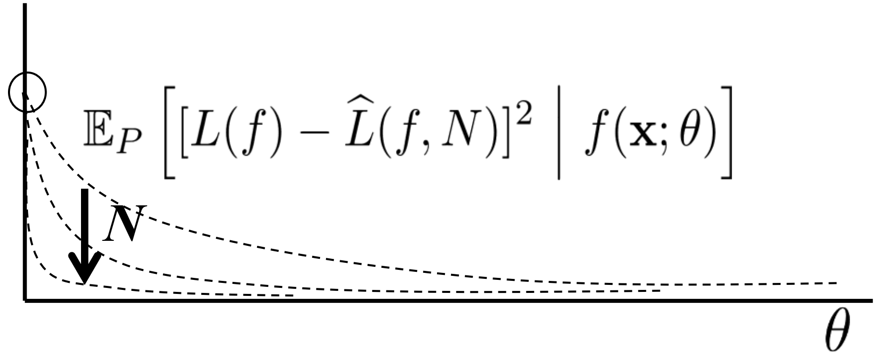
Consistent Learner

- Consider a hypothesis set $\mathcal{H}$ which satisfies

$$\mathbb{E}_P\left[[L(f)-\widehat{L}(f,N)]^2 \;\middle|\; f(\mathbf{x};\theta)\right] = \left(\frac{1}{N}\right)^{\theta}$$

$$\mathcal{H} = \{f(\mathbf{x};\theta)|\theta > 0\}$$

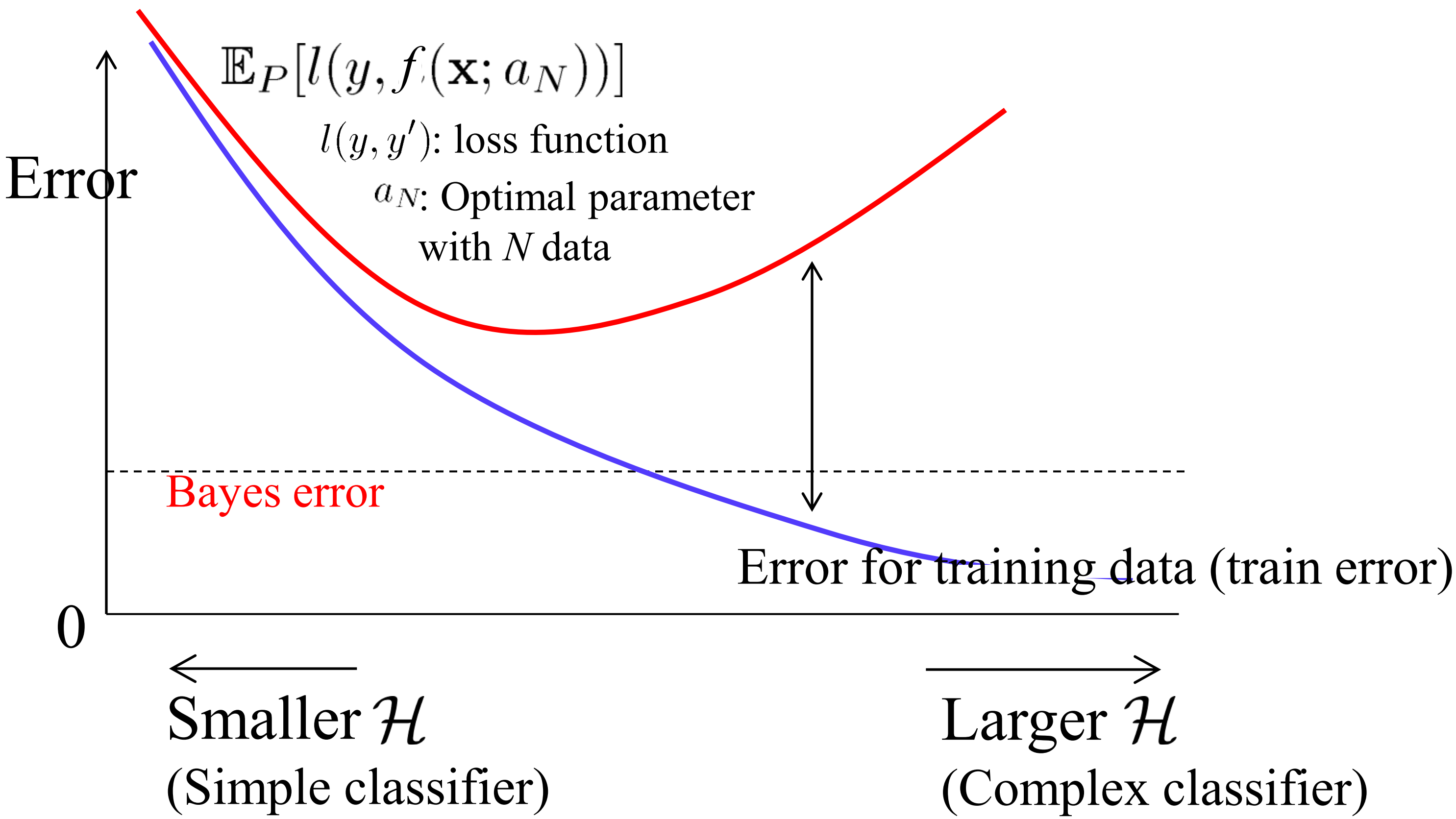
Explain that learning with $\mathcal{H}$ is not consistent though it satisfies $\widehat{L}(f) \to L(f)$.



What is the possible problem in this case?

Consistency and Bayes Error

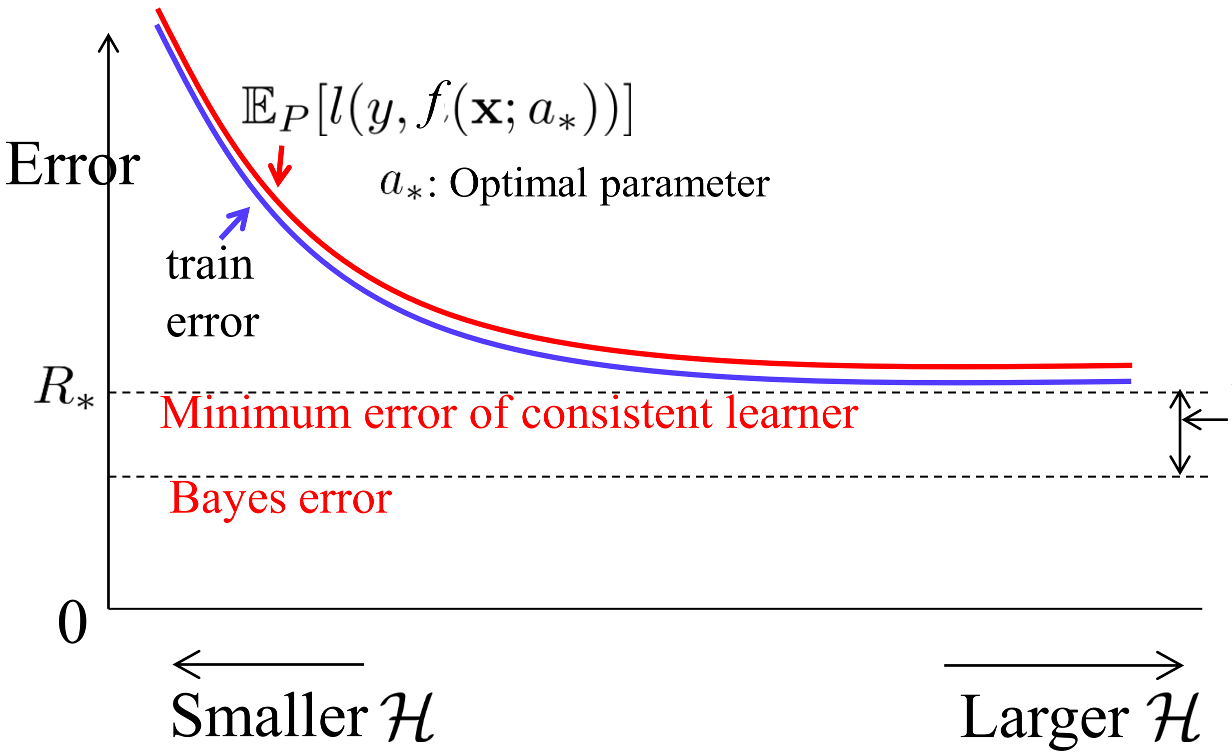
- Minimizing expected error (objective) vs. minimizing estimated error



(For example, a linear classifier with regularization)

Consistency and Bayes Error

- Consistent learner with many data

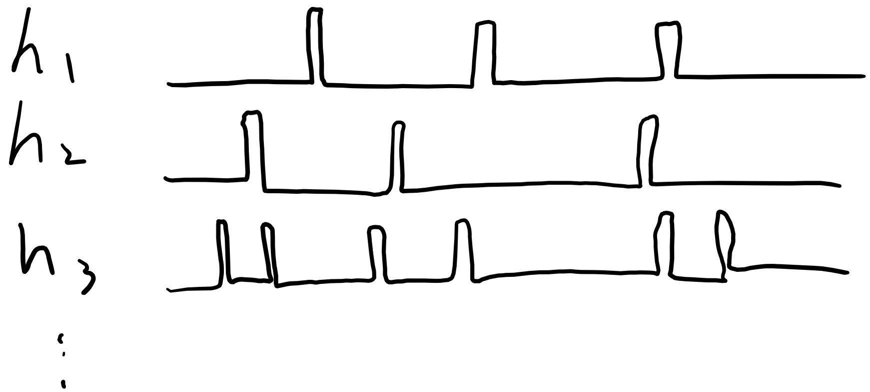


Consistency does not necessarily imply achieving the Bayes error!!!

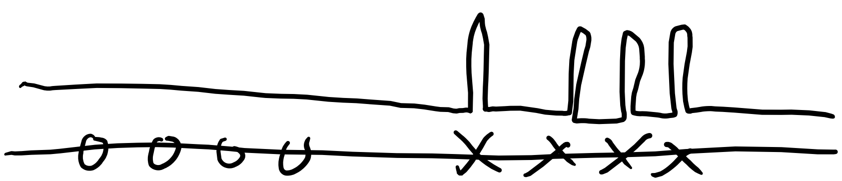
(For example, a linear classifier with regularization)

When the Model Cannot Learn (from Data)?

- Model $\mathcal{H}$



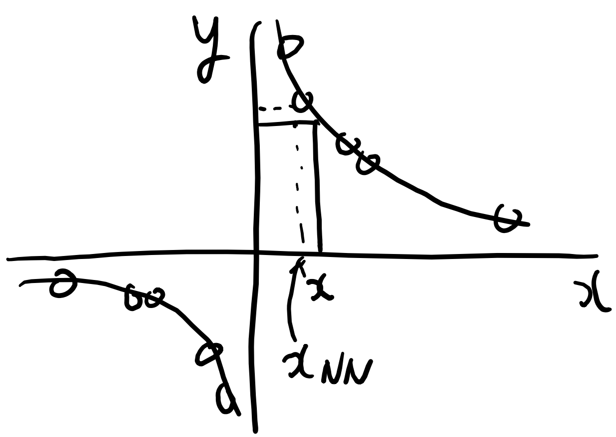
For training data (o : class 0, x : class 1)



When the Model Cannot Learn (from Data)?

- Problem

Regression $y = \frac{1}{x}$



Nearest neighbor regression

$y(x) = y(x_{NN})$

- For any x, $(y(x) - y(x_{NN}))^2 \to 0$

- $L(\mathcal{H}) = E_x (y(x) - y(x_{NN}))^2 \not\to 0$

Computations for Machine Learning

For given model $\mathcal{H} = \left\{ f(\mathbf{x};\theta) \;\middle|\; \theta \in \Theta \right\}$

- Optimization (Frequentist)
  - Find $f(\mathbf{x};\theta^*)$ s.t.

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} \left(f(\mathbf{x}_i;\theta) - y_i\right)^2 + \lambda\, \Omega(\theta)$$

- Integration (Bayesian)
  - Obtain y from

$$f(\mathbf{x};\theta)$$

$$y = \arg\max_y P(y|\mathbf{x},\mathcal{D}) = \arg\max \int P(y|\mathbf{x};\theta)p(\theta|\mathcal{D})d\theta$$

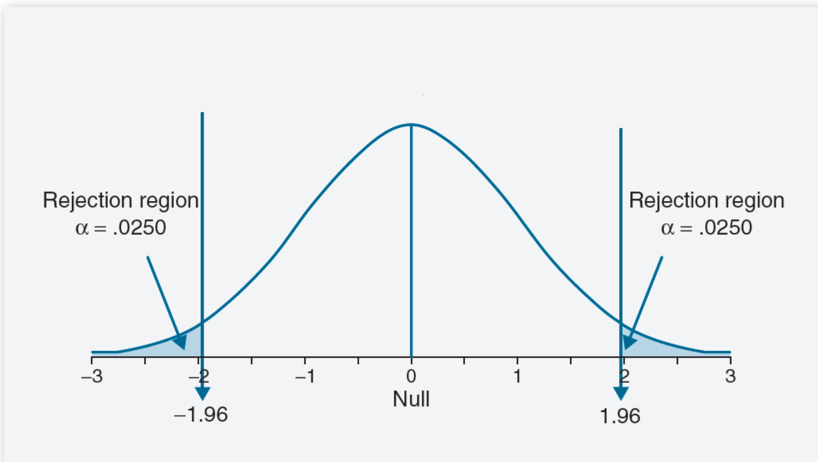
Classification

$$y = \int y\; p(y|\mathbf{x},\mathcal{D})dy = \iint y\; p(y|\mathbf{x},\theta)p(\theta|\mathcal{D})d\theta dy$$

Regression

Relationship to Traditional Statistics

- Statistics for Science
  - Hypothesis testing
  - Pursuit of Truth
- Statistical test for High-dimensional Data

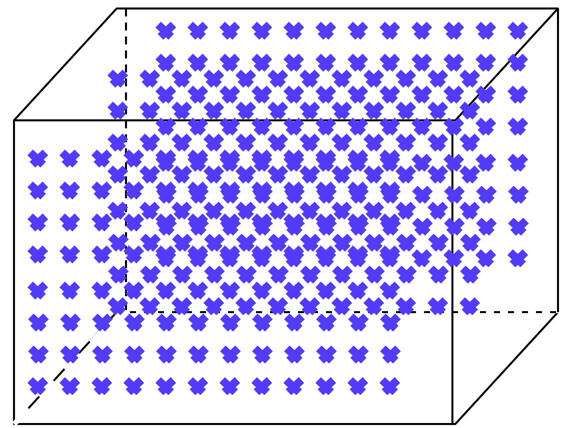


  - High-dimensional spaces usually have a large amount of discriminativity. Are the underlying densities separated? YES! (Null-hypothesis rejected!)

Curse of Dimensionality

– To achieve same density as N = *100* for *1*-variable

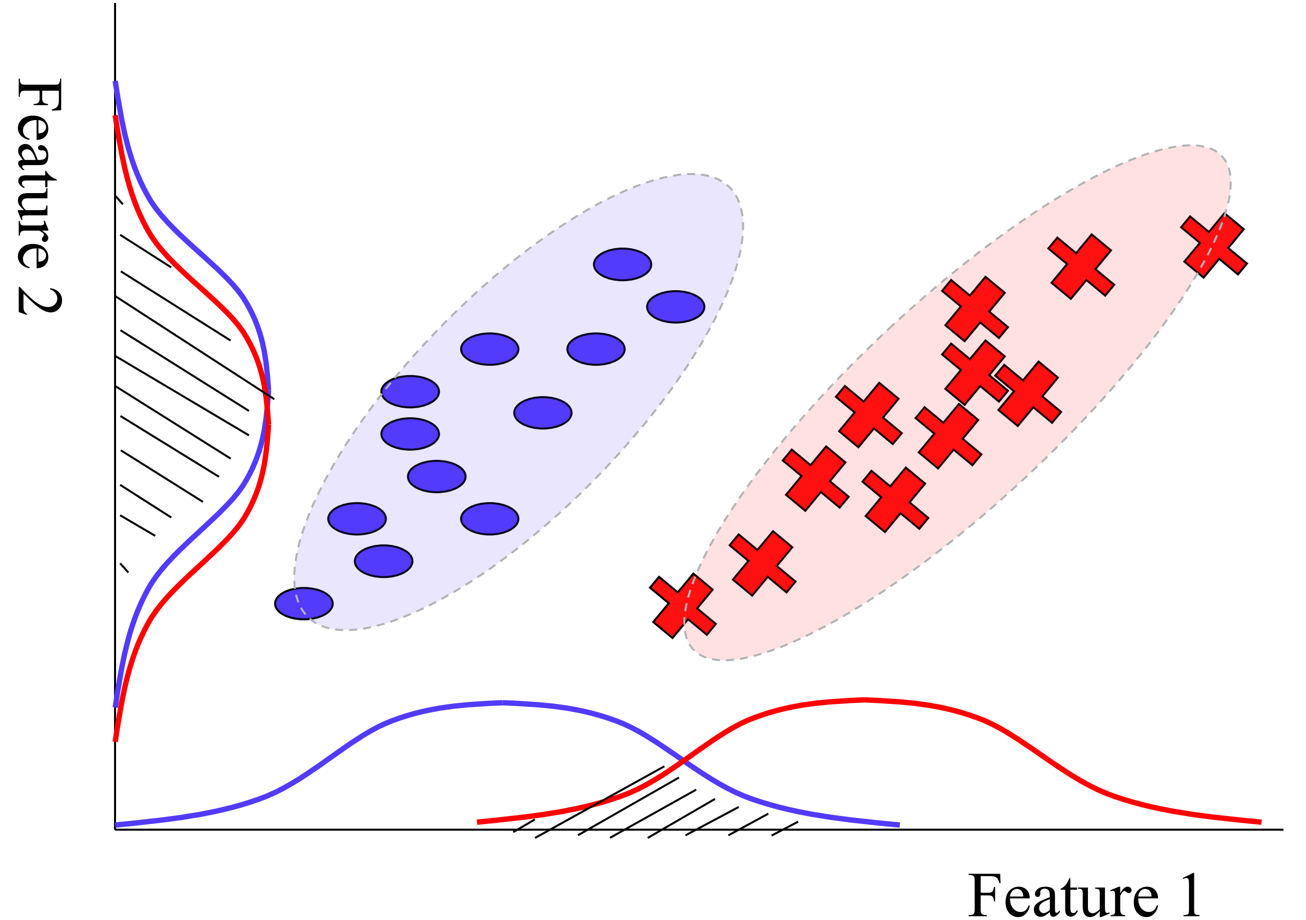
– We need N = 100^D for *D* variables



– Conversely, when we have *60,000* data for *10*-dimensional space, the density is the same as *3* data in *1*-dimensional space.

Two-Dimensional Benefits

- Feature 1 and Feature 2 have correlation




Thank you ..

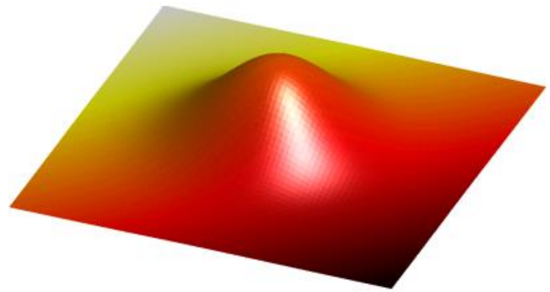
- Everything for Gaussians (?!)
- Parameter estimation for Gaussians
- Inference using Gaussians
- Gaussian Processes – Infinite dimensional Gaussians (function space view)

Gaussian Densities

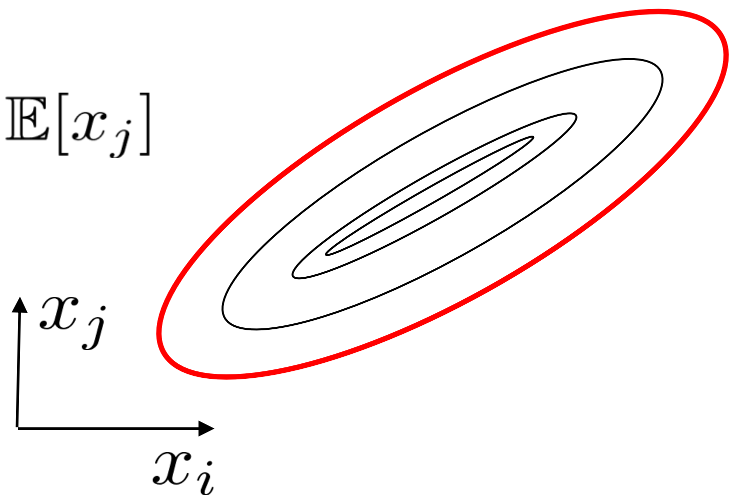
Gaussian Random Variable

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$



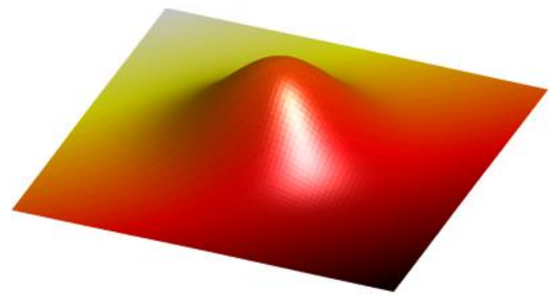
$$[\Sigma]_{ij} = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]$$



Gaussian Random Variable

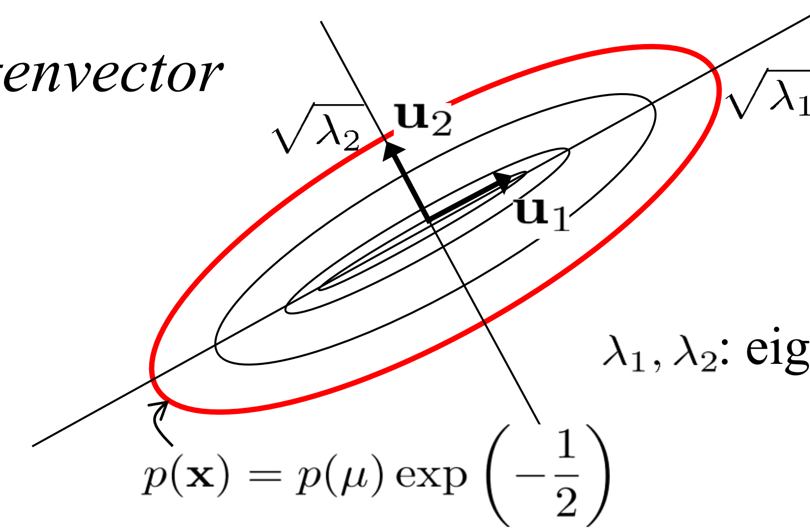
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$



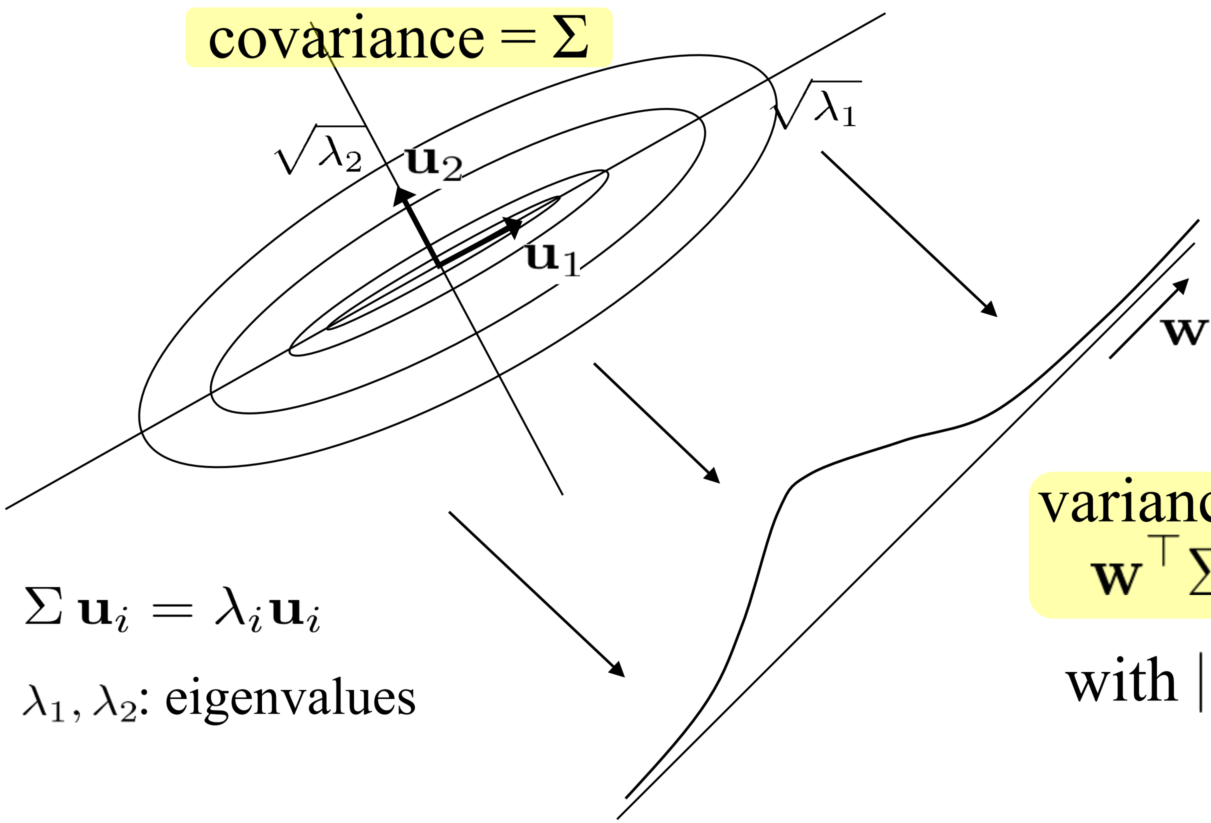
Principal axes are the eigenvector directions of Σ

$$\Sigma\,\mathbf{u}_i = \lambda_i \mathbf{u}_i$$



λ_1, λ_2: eigenvalues

Covariance Matrix and Projection



$$\Sigma\,\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

λ_1, λ_2: eigenvalues

variance = $\mathbf{w}^\top \Sigma \mathbf{w}$

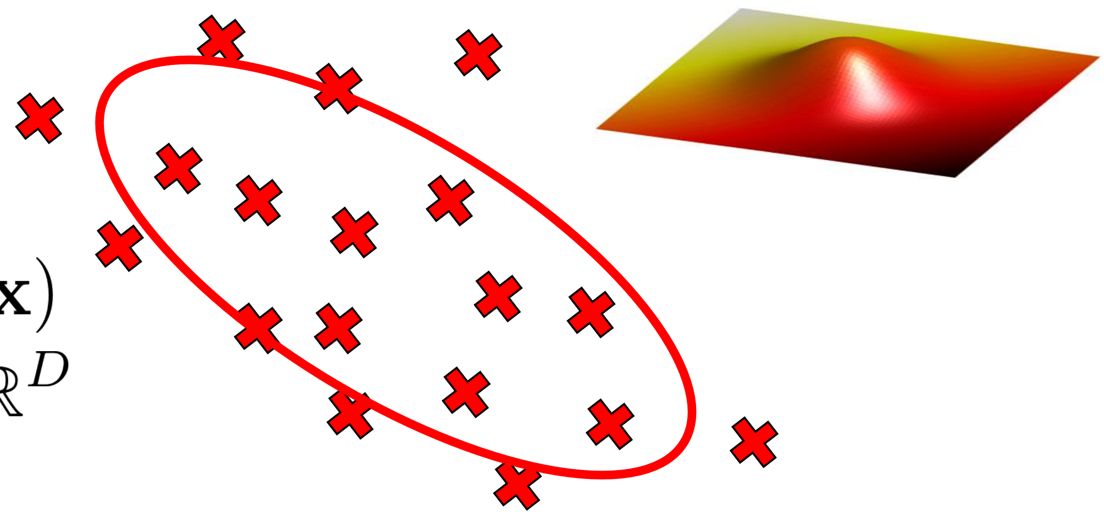
with $||\mathbf{w}||^2 = 1$

PARAMETER ESTIMATION

Motivation – Parameter Estimation

- Parameter estimation is an optimization problem

$$\mathbf{x} \sim p(\mathbf{x})$$
$$\mathbf{x} \in \mathbb{R}^D$$

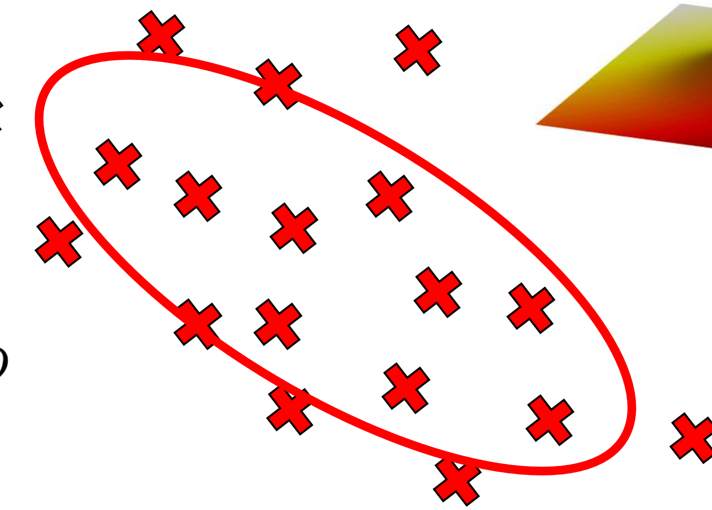


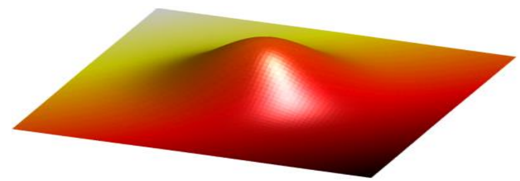
$\widehat{p}(\mathbf{x})$: estimated probability density function,
in other words, density function that fits data the most

Maximum Likelihood Estimation

- Parameter estimation is an optimization problem

$$\mathbf{x} \sim p(\mathbf{x})$$
$$\mathbf{x} \in \mathbb{R}^D$$





$$\widehat{p}(\mathbf{x}) = p(\mathbf{x}|\widehat{\mu}, \widehat{\Sigma})$$

$$\widehat{\mu}, \widehat{\Sigma} = \arg\max_{\mu,\Sigma} p(\mathbf{x}|\mu, \Sigma)$$

Maximum Likelihood for Gaussian

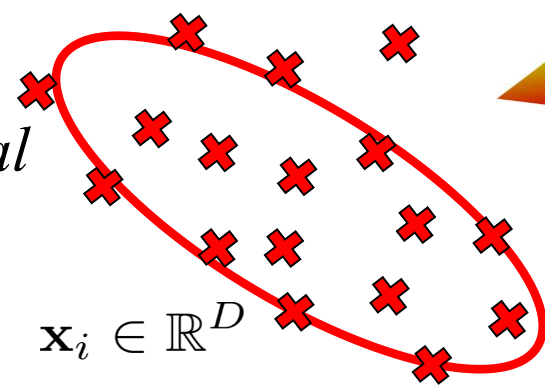
$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

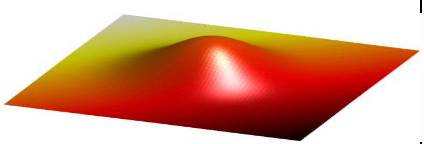
- With optimal parameters satisfying

$$\widehat{\mu}, \widehat{\Sigma} = \arg\max_{\mu,\Sigma} p(X|\mu,\Sigma) = \arg\max_{\mu,\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i|\mu,\Sigma)$$

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^\top$$

Empirical mean and empirical covariance are the maximum likelihood solutions.





$\widehat{\mu}, \widehat{\Sigma}$

$\mathbf{x}_i \in \mathbb{R}^D$

Maximum Likelihood for Gaussian

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

$$\boxed{\nabla_\theta \ln p(X|\theta) = \vec{0} \qquad \theta = \mu, \Sigma}$$

$$\frac{\partial \ln p(X|\mu,\Sigma)}{\partial \mu} = 0 \Longrightarrow \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

$$\frac{\partial \ln p(X|\mu,\Sigma)}{\partial \Sigma} = 0 \Longrightarrow \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^\top$$

Maximum A Posteriori (MAP) Estimation

- MAP estimation

$$\theta^* = \arg\max_\theta p(\theta|X) \qquad \text{cf) } \theta^* = \arg\max_\theta p(X|\theta)$$

- Likelihood (Model): $p(\mathbf{x}|\theta)$
- Prior: $p(\theta)$
- Bayes rule:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

Maximum A Posteriori (MAP) Estimation for Gaussian

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

$$\widehat{\mu} = \arg\max_{\mu} p(\mu|X) = \arg\max_{\mu} \prod_{i=1}^{N} p(\mu|x_i)$$

- Let the prior

$$p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

- The posterior can be calculated using

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{i=1}^{N} p(x_i|\mu)p(\mu) \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

Maximum A Posteriori (MAP) Estimation for Gaussian

$$\prod_{i=1}^{N} p(x_i|\mu)p(\mu) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right]$$

$$\cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

$$\propto \quad \exp\left(-\frac{1}{2}\left(\sum \frac{(x_i-\mu)^2}{\sigma^2} + \frac{\mu-\mu_0}{\sigma_0^2}\right)\right)$$

$$\propto \quad \exp\left(-\frac{1}{2}\left(\mu^2\left[\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right] - 2\mu\left[\frac{1}{\sigma^2}\sum x_i + \frac{\mu_0}{\sigma_0}\right]\right)\right)$$

$$\propto \exp(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2)$$

Maximum A Posteriori (MAP) Estimation for Gaussian

- Posterior density

$$\propto \quad \exp\left(-\frac{1}{2}\left(\mu^2\left[\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right]-2\mu\left[\frac{1}{\sigma^2}\underbrace{\sum x_i}_{=N\widehat{\mu}_{ML}}+\frac{\mu_0}{\sigma_0}\right]\right)\right)$$

  - Caution: Posterior of μ, not the density function of x

- MAP of μ = Mean of μ = μ_n

$$\mu_n = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\widehat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

MLE vs. MAP

- For Gaussian
  - When N is just a few (say N = 5),

$$\sigma_0^2 = 5, \sigma^2 = 3$$

$$\mu_n = \underbrace{\frac{25}{5 \cdot 5 + 3} \widehat{\mu}_{ML}}_{\text{Dominant}} + \frac{3}{5 \cdot 5 + 3} \mu_0$$

$$\sigma_n = \frac{5 \cdot 3}{25 + 3} \fallingdotseq 0.54$$

MLE vs. MAP

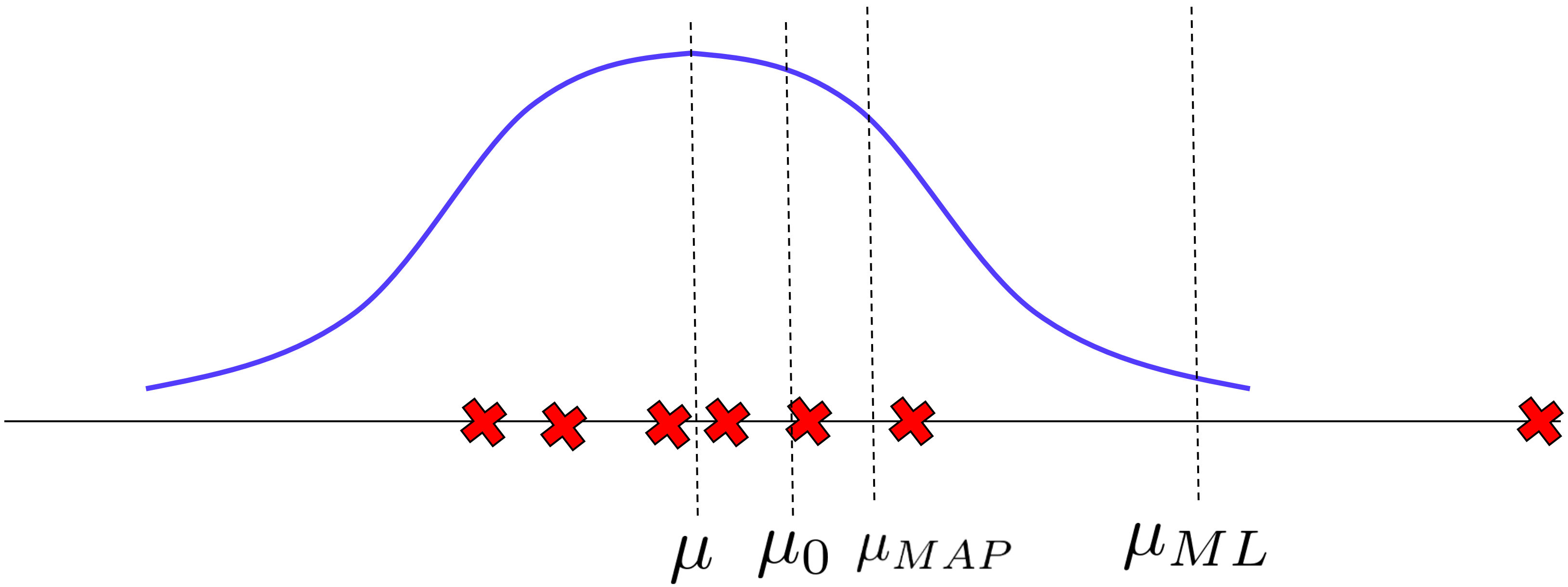
- For Gaussian
  - When we have a few outliers

$$\sigma_0^2 = 5, \sigma^2 = 100$$

$$\mu_n = \frac{25}{5 \cdot 5 + 100} \widehat{\mu}_{ML} + \underbrace{\frac{100}{5 \cdot 5 + 100} \mu_0}_{\text{Dominant (learn from } \mu_0\text{)}}$$

$$\sigma_n = \frac{5 \cdot 100}{25 + 100} \fallingdotseq 4$$

MLE vs. MAP



Bayesian Integration

- The final standard method of prediction is to use Bayesian inference instead of estimating the parameter point.
  - Do not insert $\widehat{\mu}_{MAP}$ directly, but marginalize.

$$p(x|X) = \int p(x|\mu)p(\mu|X)d\mu$$

$$= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{1}{2\sigma_n}(\mu-\mu_n)^2\right) d\mu$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_n^2)}} \exp\left(-\frac{1}{2(\sigma^2+\sigma_n^2)}(x-\mu)^2\right)$$

$$= \boxed{\mathcal{N}(\mu_n, \sigma^2 + \underline{\sigma_n^2})}$$

Uncertainty for prediction

Inference using Gaussians

Decomposition for Inference

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$= C\exp\left(-\frac{1}{2}\left(\mathbf{x}_a - \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)^\top \left(\Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba}\right)^{-1}\left(\mathbf{x}_a - \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)\right.$$
$$\left. -\frac{1}{2}\left(\mathbf{x}_b-\mu_b\right)^\top \Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)$$

$$\Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba}$$

$$\mu_{a|b} = \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b-\mu_b)$$

$$= C\exp\left(-\frac{1}{2}(\mathbf{x}_a-\mu_{a|b})^\top \Sigma_{a|b}^{-1}(\mathbf{x}_a-\mu_{a|b})\right.$$
$$\left. -\frac{1}{2}(\mathbf{x}_b-\mu_b)^\top \Sigma_b^{-1}(\mathbf{x}_b-\mu_b)\right)$$

Decomposition for Inference

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

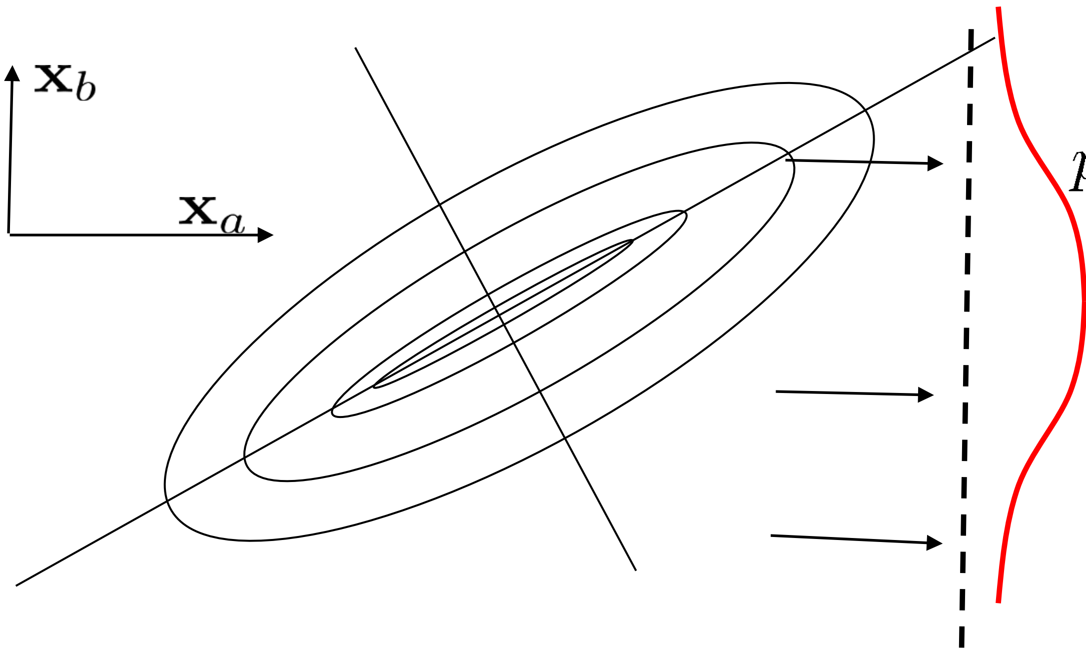
$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1} (\mathbf{x}-\mu) \right) \\ &= C_1 \exp\left(-\frac{1}{2}(\mathbf{x}_a - \mu_{a|b}(\mathbf{x}_b))^\top \Sigma_{a|b}^{-1} (\mathbf{x}_a - \mu_{a|b}(\mathbf{x}_b)) \right) \cdot \\ &\qquad C_2 \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mu_b)^\top \Sigma_b^{-1} (\mathbf{x}_b - \mu_b) \right) \end{aligned}$$

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

Gaussian Random Variable – Marginal

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

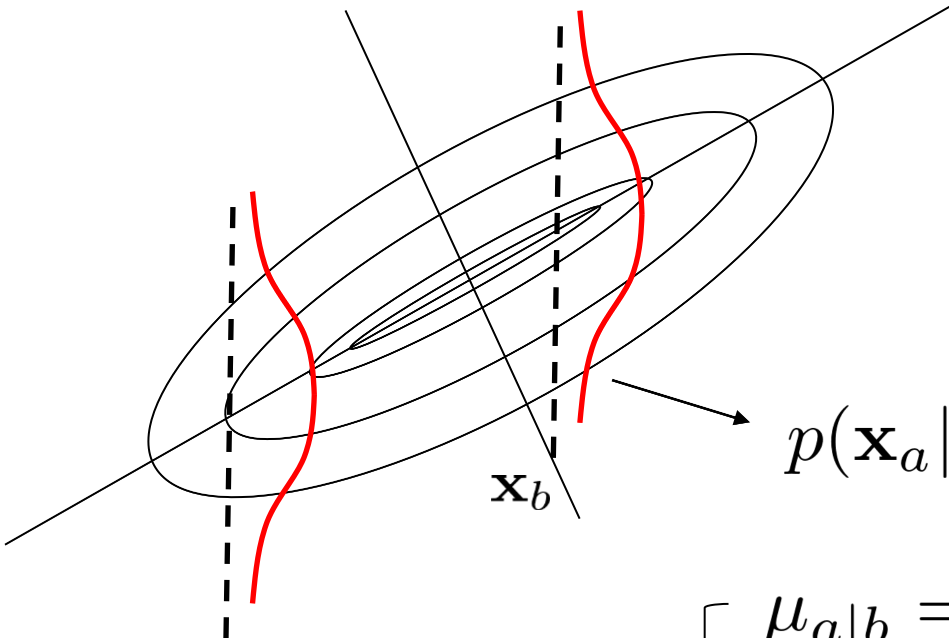
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1} (\mathbf{x}-\mu) \right)$$



$$\begin{aligned} p(\mathbf{x}_b) &= \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_a \\ &= \int p(\mathbf{x}_a | \mathbf{x}_b) p(\mathbf{x}_b) d\mathbf{x}_a \\ &= \mathcal{N}(\mu_b, \Sigma_b) \end{aligned}$$

Gaussian Random Variable – Conditional

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$



$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix}$$

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$$

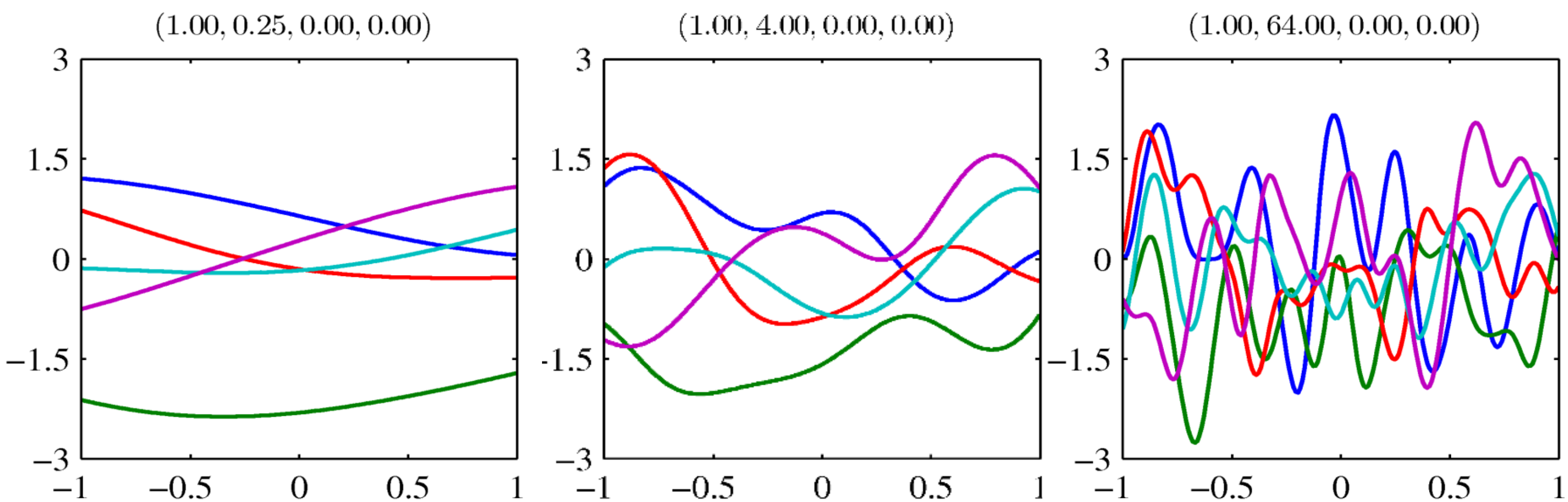
$$\begin{cases} \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b - \mu_b) \\ \Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba} \end{cases}$$

Gaussian Processes – Function Space View

$$y(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

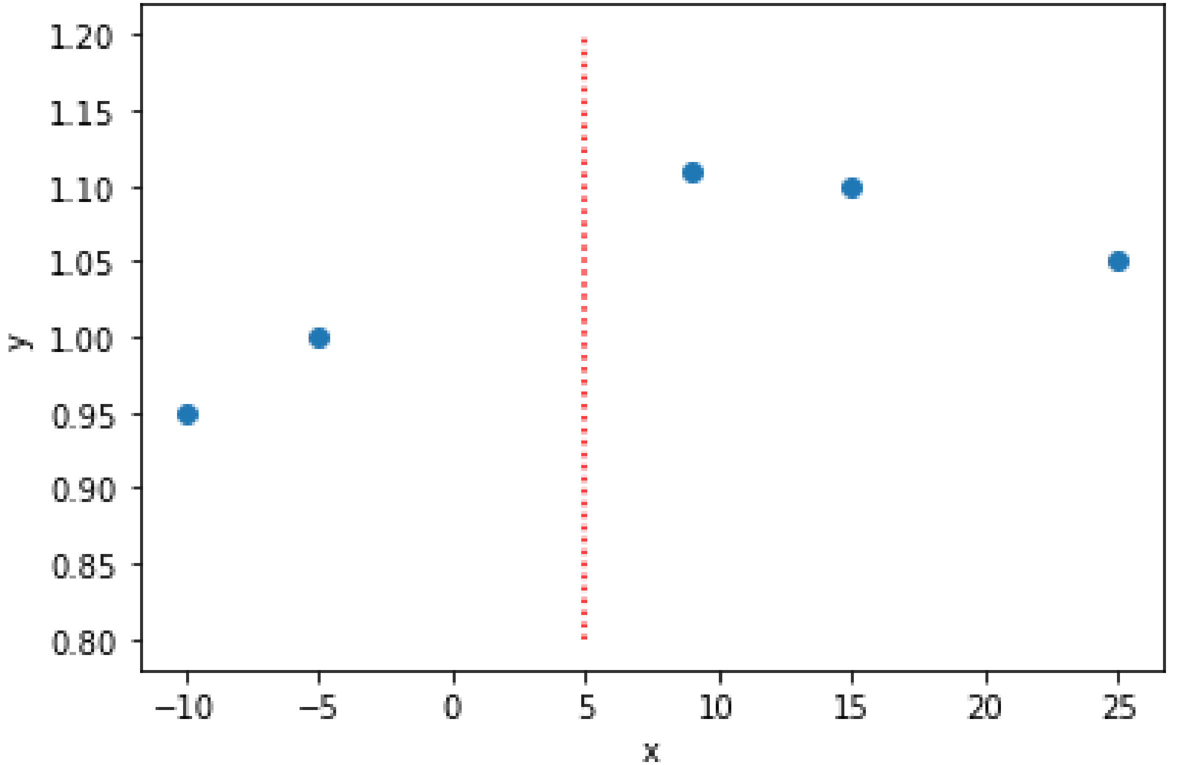
$$m(\mathbf{x}) = \mathbb{E}[y(\mathbf{x})] = 0$$

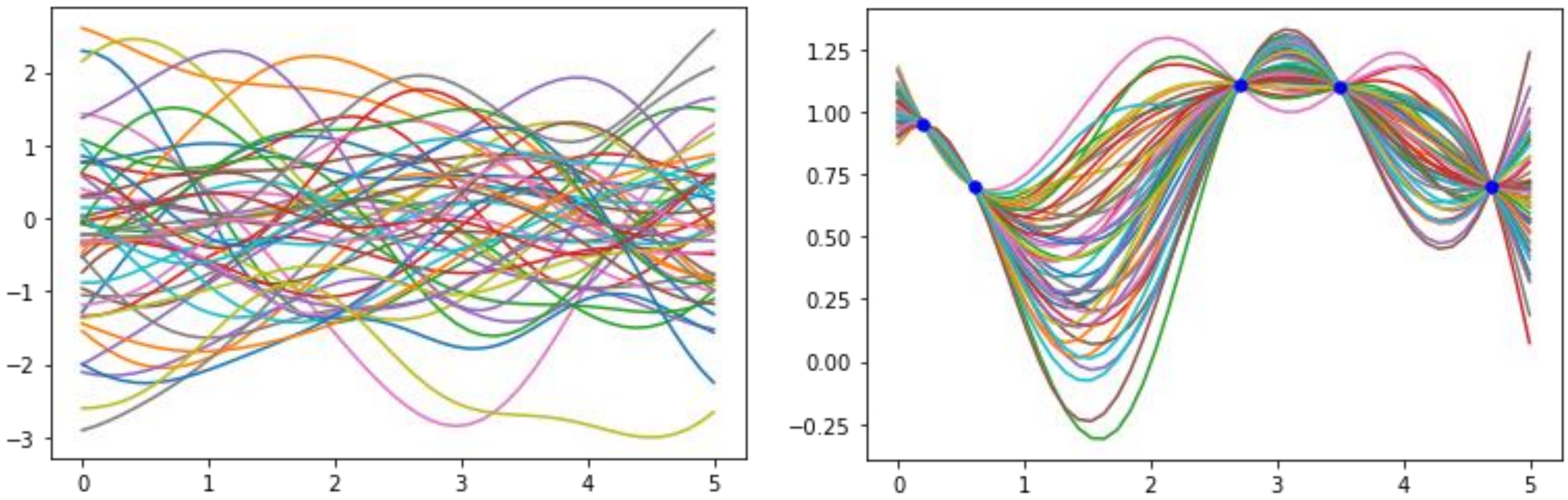
$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(y(\mathbf{x}) - m(\mathbf{x}))(y(\mathbf{x}') - m(\mathbf{x}'))] \\ &= \theta_1 \exp\left\{-\frac{\theta_2}{2}||\mathbf{x} - \mathbf{x}'||^2\right\} + \theta_3 + \theta_4 \mathbf{x}^\top \mathbf{x}' \end{aligned}$$

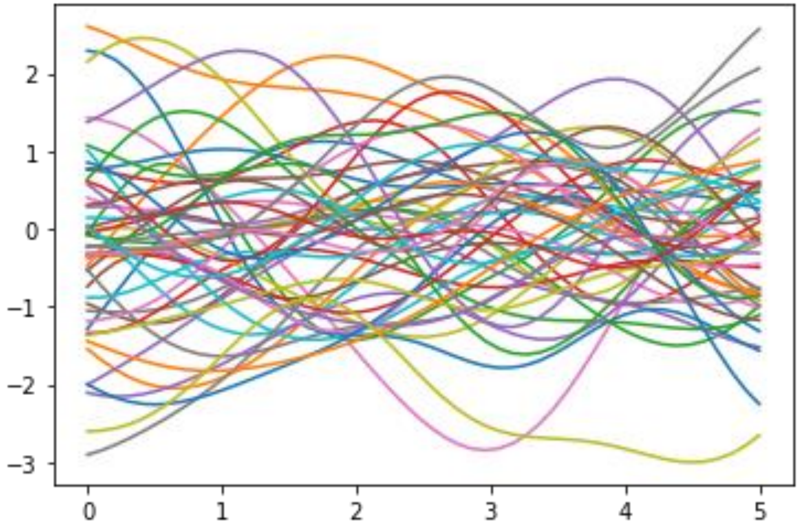


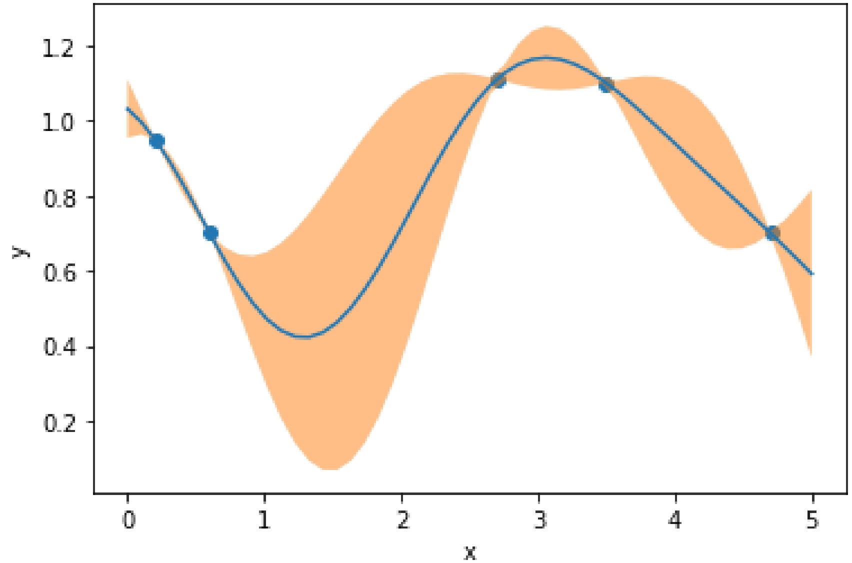
C. Bishop (2007) *Pattern Recognition and Machine Learning*, Springer

Regression




2
1
0
−1
−2
−3
0
1
2
3
4
5
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
1
2
3
4
5





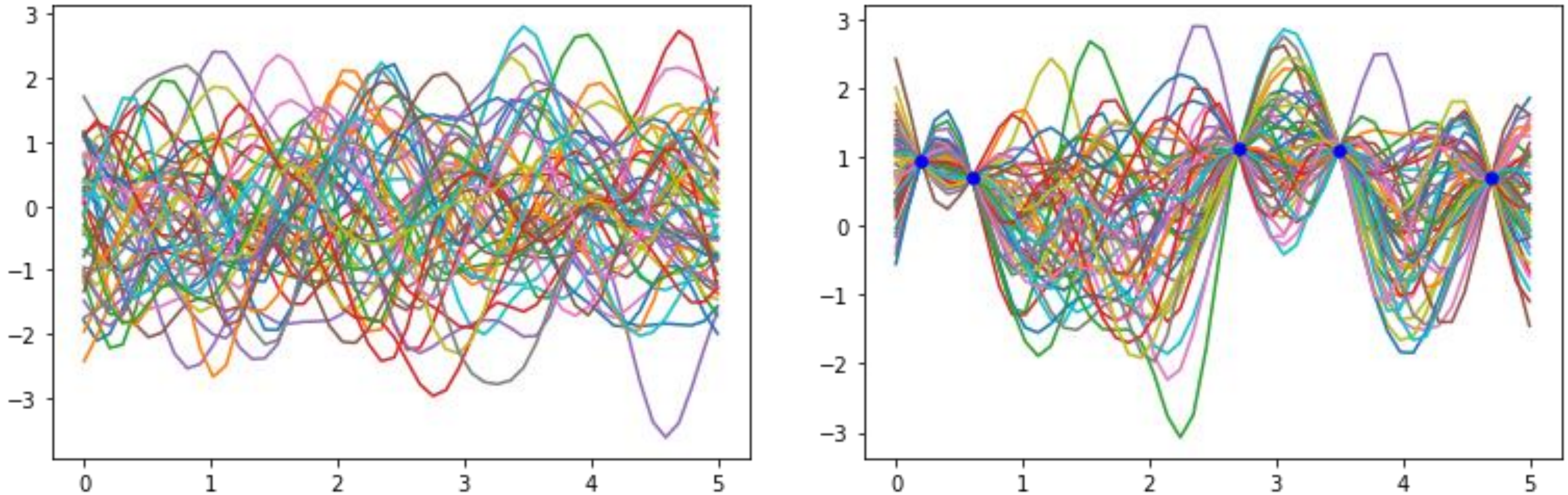
$$m(x) = \mathbf{k}^\top K^{-1}\mathbf{y}$$

$$\sigma^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

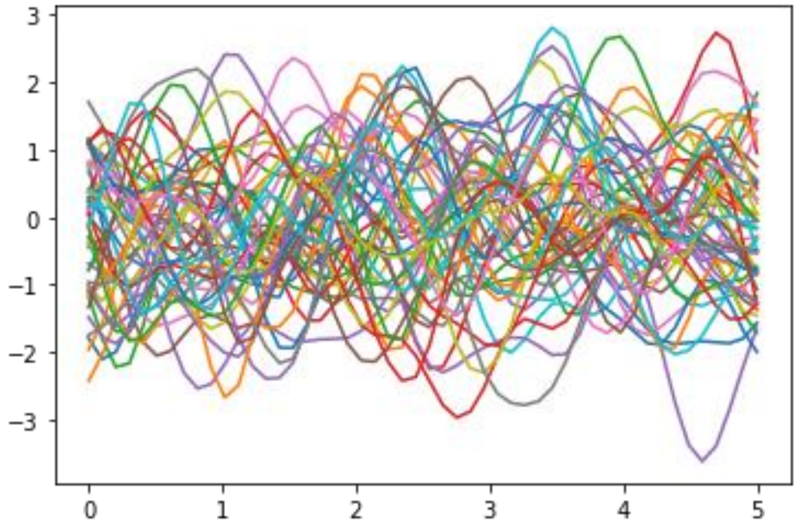
$$[\mathbf{k}]_i = k(\mathbf{x},\mathbf{x}_i)$$

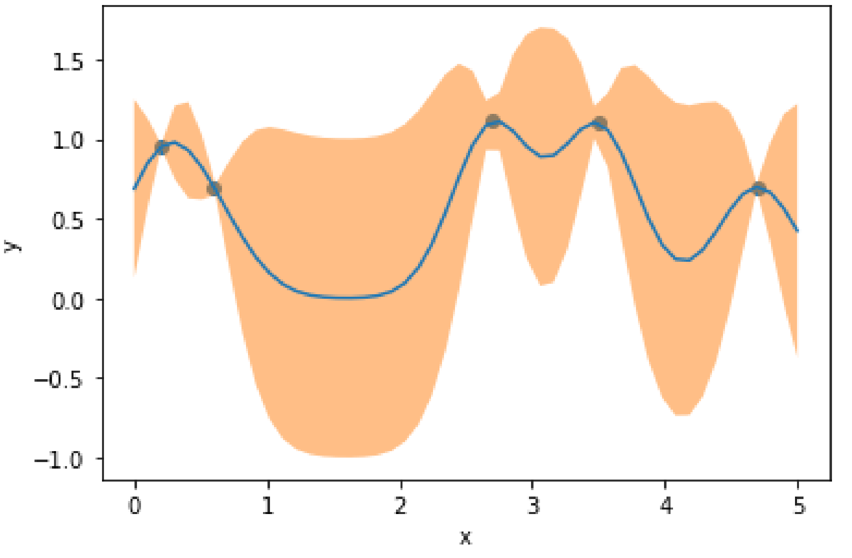
$$[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$$

$$[\mathbf{y}]_i = y(\mathbf{x}_i)$$



3
2
1
0
−1
−2
−3
0
1
2
3
4
5





$$m(x) = \mathbf{k}^\top K^{-1}\mathbf{y}$$

$$\sigma^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

$$[\mathbf{k}]_i = k(\mathbf{x},\mathbf{x}_i)$$

$$[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$$

$$[\mathbf{y}]_i = y(\mathbf{x}_i)$$

Bayesian linear regression

Review – Two Learning Paradigms

Data:

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \qquad \mathbf{x} \in \mathbb{R}^D, y_i \in \{1, \ldots, C\} \text{ or } y_i \in \mathbb{R}$$

Model:

$$f(\mathbf{x};\theta) \in \mathcal{H} \quad \text{or} \quad p(\mathcal{D}|\theta) \in \mathcal{H}$$

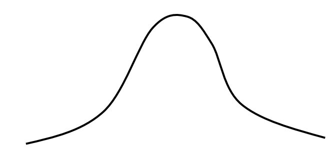
- 1) Choose the best fit to the data in terms of $L(y, f)$

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i))$$

Prediction: $y = f(\mathbf{x};\theta^*)$

- 2) Choose the best guess with likelihood

Likelihood: $p(\mathcal{D}|\theta)$ Prior: $p(\theta)$



Posterior: $p(\theta|\mathcal{D}) = \dfrac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$

Prediction: $p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x};\theta)p(\theta|\mathcal{D})d\theta$