

# KSBI-BIML 2026

Bioinformatics & Machine Learning(BIML)  
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)



## An Introduction to Probabilistic Modeling

노영균 \_ 한양대학교



**KSBI**  
KOREAN SOCIETY FOR  
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로  
제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우  
발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

# KSBI-BIML 2026

## Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics & Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의를 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

## An Introduction to Probabilistic Modeling

본 강의에서는 학습과 예측을 위한 확률 모델링에 대한 기초와 응용 방법에 대해 설명한다. 기계 학습의 여러 기본 개념이 왜 확률 모델링과 관계를 가지는지 설명하고, 확률 모델을 이용하기 위한 기초적인 학습 방법과 추론 방법에 대해 소개한다.

다양한 학습 방법이 사용 가능한 데이터 수에 어떤 영향을 받는지 설명하고, 다양한 확률 모델링의 정보 이론적 해석에 대해 설명한다.

고차원 확률 모델링의 이슈와 그 해결책에 대한 개념 파악을 주요한 목표로 한다. 시간이 허락하면 정보이론값 추정에 대해 간단히 소개한다.

\* 강의 난이도: 초급

\* 강의: 노영균 교수 (한양대학교 컴퓨터소프트웨어학부)

## Curriculum Vitae

**Speaker Name: Yung-Kyun Noh, Ph.D.**



### ► Personal Info

Name Yung-Kyun Noh  
Title Assistant Professor  
Affiliation Hanyang University

### ► Contact Information

Email nohyung@hanyang.ac.kr

---

### Research Interest

Machine Learning, Nonparametric methods, Information theory

### Educational Experience

2011 Ph.D. in Interdisciplinary Program in Cognitive Science, Seoul National University, Korea  
1998 B.S. in Physics, POSTECH, Korea

### Professional Experience

2019- Assistant Professor, Dept. of Computer Science, Hanyang University, USA  
2019- Associate Member, Korea Institute for Advanced Study (KIAS), Korea  
2020- Visiting Scientist, Gastroenterology, Mayo Clinic at Rochester, MN, USA  
2018- Visiting Scientist, RIKEN Center for Advanced Intelligence Project (API), Japan  
2015-2018 BK Assistant Professor, Dept. of Mechanical and Aerospace Engineering, Seoul National University, Korea  
2013-2014 Research Assistant Professor, Dept. of Computer Science, KAIST, Korea  
2011-2013 Postdoctoral fellow, Dept. of Mechanical and Aerospace Engineering, Seoul National University, Korea  
2007-2012 Visiting Researcher, Dept. of Electrical and Systems Engineering, University of Pennsylvania, PA, USA

### Selected Publications (5 maximum)

1. Noh, Y.K., Park, J., Choi, B. G., Kim, K.-E., and Rha, S.W. (2019) A Machine Learning-Based Approach for the Prediction of Acute Coronary Syndrome Requiring Revascularization, *Journal of Medical Systems*, 43(8), Article 253
2. Ganguly, S., Ryu, J., Kim, Y.H., Noh, Y.K., Lee, D.D. (2018) Nearest neighbor density functional estimation based on inverse Laplace transform, *arXiv:1805.08342*
3. Noh, Y.K., Hamm, J.H., Park, F.C., Zhang, B.T., and Lee, D.D. (2018) Fluid Dynamic Models for Bhattacharyya-based Discriminant Analysis, *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 40(1):92-105
4. Noh, Y.K., Zhang, B.T., and Lee, D.D. (2018), Generative Local Metric Learning for Nearest Neighbor Classification, *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 40(1):106-118
5. Noh, Y.K., Sugiyama, M., Kim, K.E., Park, F.C., and Lee, D.D. (2017), Generative Local Metric Learning for Kernel Regression, *Advances in Neural Information Processing Systems* 30

# KSBi-BIML 2021

An Introduction to Probabilistic Modeling  
Yung-Kyun Noh (Hanyang University)

Organized by



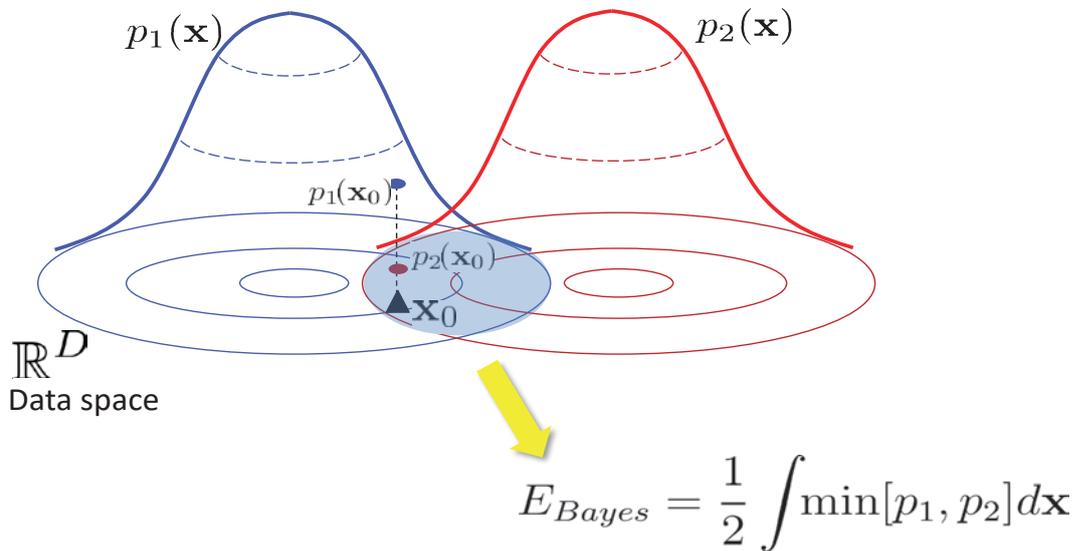
<http://aais.hanyang.ac.kr>

## Contents

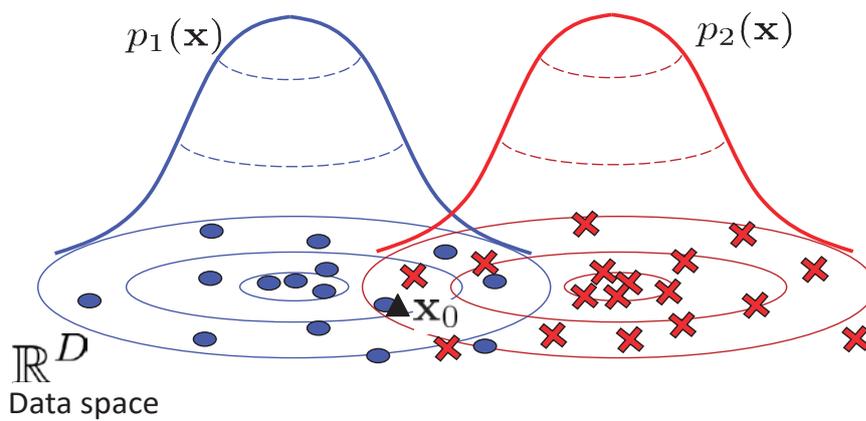
- Properties of probability models and probability density models
- Parameter estimation
- Inference

# Probabilistic Assumption and Bayes Classification

- Bayes Error



# Probabilistic Assumption and Bayes Classification

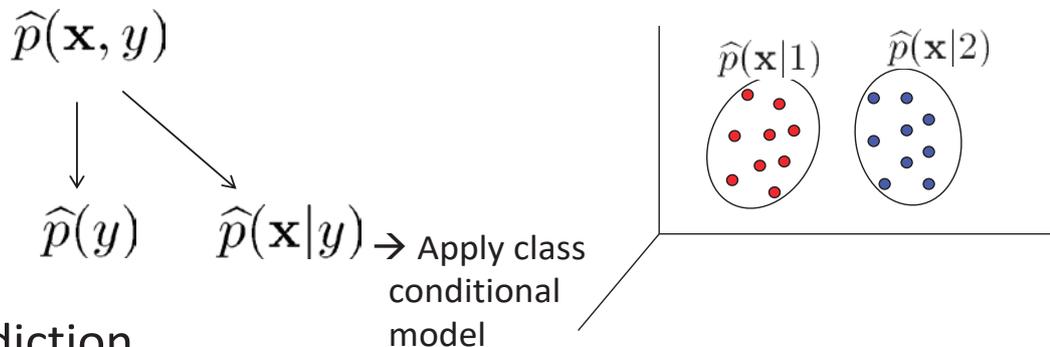


$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$



# Generative vs. Discriminative (1/2)

- Generative Learning
  - Interested in joint probability



- Prediction
  - Bayes Rule

$$h(\mathbf{x}) = \hat{p}(y|\mathbf{x}) = \frac{\hat{p}(y)\hat{p}(\mathbf{x}|y)}{\hat{p}(\mathbf{x})}$$

# Generative vs. Discriminative (2/2)

- Discriminative learning
  - “NOT” interested in joint probability
- Conditional probability learning

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

- Learn prediction function by minimizing the empirical loss function

$$h(\mathbf{x}) = \arg \min_{h \in \mathcal{H}} \hat{\epsilon}(\mathcal{D}, h)$$

V. N. Vapnik (1998) *Statistical learning theory*, John Wiley & Sons  
Also refer to NIPS 2009 workshop -- Generative / Discriminative Interface

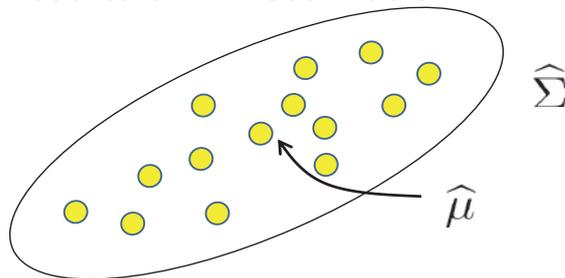
# Learn Density Function from Data

- Make a model with learning parameters
  - Statistically obtain parameter values using ML or MAP estimation
  - In Gaussian,

$$\widehat{mean} = \frac{1}{N} \sum_i \mathbf{x}_i \equiv \widehat{\mu}$$

$$\widehat{covariance} = \frac{1}{N} \sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \equiv \widehat{\Sigma}$$

are the results of ML estimation.



## For Two Gaussian Data (1/2)

$$\widehat{p}_1(\mathbf{x}) = \mathcal{N}(\widehat{\mu}_1, \widehat{\Sigma}_1)$$

$$\widehat{p}_2(\mathbf{x}) \sim \mathcal{N}(\widehat{\mu}_2, \widehat{\Sigma}_2)$$

$$\widehat{p}_1(\mathbf{x}) \geq \widehat{p}_2(\mathbf{x}) \iff \frac{\widehat{p}_2(\mathbf{x})}{\widehat{p}_1(\mathbf{x})} \geq 1$$

$$\frac{\frac{1}{\sqrt{2\pi^D} |\widehat{\Sigma}_2|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \widehat{\mu}_2)^T \widehat{\Sigma}_2^{-1} (\mathbf{x} - \widehat{\mu}_2)\right)}{\frac{1}{\sqrt{2\pi^D} |\widehat{\Sigma}_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \widehat{\mu}_1)^T \widehat{\Sigma}_1^{-1} (\mathbf{x} - \widehat{\mu}_1)\right)} \geq 1$$

## For Two Gaussian Data (2/2)

- With a Homoscedastic Assumption

$$\hat{\Sigma}_1 = \hat{\Sigma}_2 \equiv \hat{\Sigma}$$

$$\exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\mathbf{x} - \hat{\mu}_2) + \frac{1}{2}(\mathbf{x} - \hat{\mu}_1)^T \hat{\Sigma}^{-1}(\mathbf{x} - \hat{\mu}_1)\right) \geq 1$$

The problem reduces to

$$\exp(\mathbf{w}^T \mathbf{x} - b) \geq 1$$

$$\mathbf{w} = \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

$$b = \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 - \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1$$

→ Fisher Discriminant Analysis



## In Terms of the Posterior

$$\begin{aligned} p(y = 1 | \mathbf{x}, \mathbf{w}, b) &= \frac{p_1}{p_1 + p_2} \\ &= \frac{1}{1 + p_2/p_1} = \frac{1}{1 + \exp(\mathbf{w}^T \mathbf{x} - b)} \end{aligned}$$

$$\begin{aligned} p(y = 2 | \mathbf{x}, \mathbf{w}, b) &= 1 - p(y = 1 | \mathbf{x}) \\ &= \frac{\exp(\mathbf{w}^T \mathbf{x} - b)}{1 + \exp(\mathbf{w}^T \mathbf{x} - b)} \end{aligned}$$



# Logistic Regression

- Starts from the posterior

$$p(y = 1|\mathbf{x}, \mathbf{w}, b) = \frac{1}{1 + \exp(\mathbf{w}^T \mathbf{x} - b)}$$

$$p(y = 2|\mathbf{x}, \mathbf{w}, b) = \frac{\exp(\mathbf{w}^T \mathbf{x} - b)}{1 + \exp(\mathbf{w}^T \mathbf{x} - b)}$$

$$\mathbf{w}, b = \arg \max_{\mathbf{w}, b} \ln p(\mathbf{y}|X, \mathbf{w}, b)$$
$$\sum_n \mathbb{I}(y_n = 1) \ln p(y_n = 1|\mathbf{x}_n, \mathbf{w}, b)$$
$$+ \mathbb{I}(y_n = 2) \ln p(y_n = 2|\mathbf{x}_n, \mathbf{w}, b)$$

Use gradient ascent → *Local* maxima



# FDA and Logistic Regression

- Have the same discriminative form (Linear Classifier)

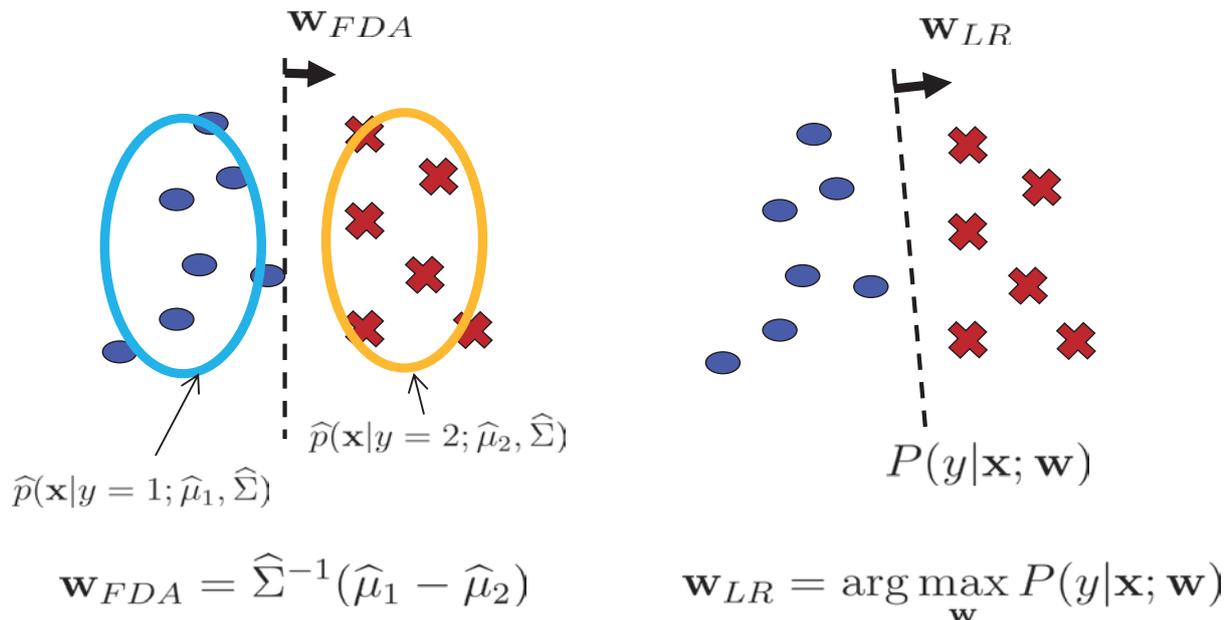
$$\mathbf{w}^T \mathbf{x} - b \geq 0$$

- FDA solution: Bayes classifier with class-conditional model
- Logistic regression: Discriminative adaptation of a discriminative function
- Question: Are the results the same or not?



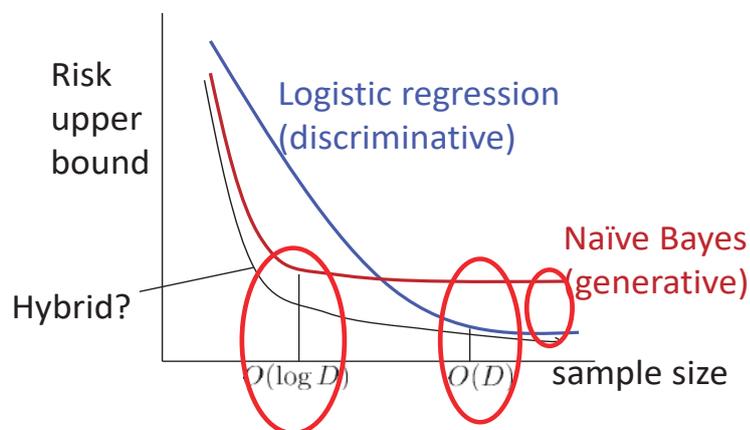
# FDA and Logistic Regression

- Are the results the same or not?



# Comparative Study (1/2)

- Generative & Discriminative Pair
  - Same number of parameters, same form of  $h(x)$



S. Lacoste-Julien et al. (2009) The generative and discriminative learning interface, *NIPS Workshop*  
 A. Y. Ng & M. I. Jordan (2001) On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes, *NIPS*

# Comparative Study (2/2)

- Discriminative analog of naïve Bayes is logistic regression
- The error  $err(f_{Disc}(\mathbf{x}))$  converges to  $err(f_{Disc,\infty}(\mathbf{x}))$ , and  $err(f_{Disc,\infty}(\mathbf{x}))$  is no worse than linear classifier picked by naïve Bayes.
- With  $O(\log D)$  samples, the parameters of  $f_{Gen}$  are close to those of  $f_{Gen,\infty}$  uniformly.
- The parameter convergence implies  $err(f_{Gen}(\mathbf{x}))$  approaches  $err(f_{Gen,\infty}(\mathbf{x}))$ .



# Keywords

- Probability / Probability density
- Conditional probability (density)

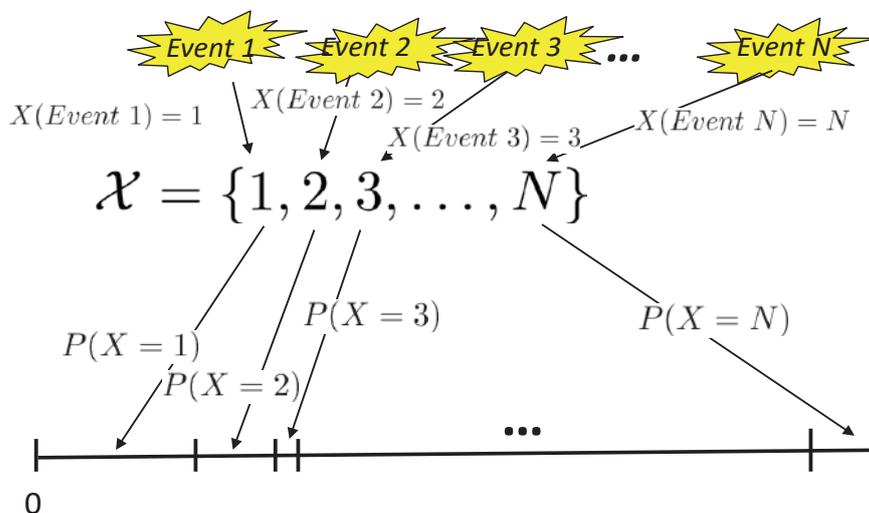
$$p(\mathbf{x}_2|\mathbf{x}_1) \quad P(y|\mathbf{x})$$
$$\mathbf{x}_1 \in \mathbb{R}^{D_1}, \mathbf{x}_2 \in \mathbb{R}^{D_2}, \mathbf{x} \in \mathbb{R}^D, y \in \{1, 2\}$$

- Marginal probability (density)
- Joint probability (density)
- Inference and classification

# Probability

$$P(X) : \mathcal{X} \rightarrow [0, 1]$$

- Mapping from a random variable to a number



# Probability

$X$ : random variable     $X_1$ : set of outputs of random variables

$$P(X_1) \equiv P(X \in X_1)$$

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

$$X_1 = \{1, 2, 3, 4\}, \quad X_2 = \{3, 4, 5\}$$

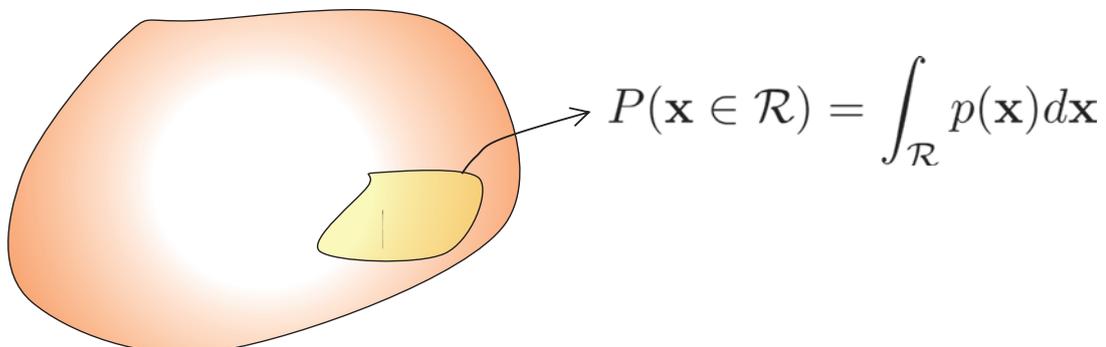
$$P(1, 2, 3, 4, 5) = P(1, 2, 3, 4) + P(3, 4, 5) - P(3, 4)$$

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) \text{ if } X_1 \cap X_2 = \phi$$



# Probability and Probability Density

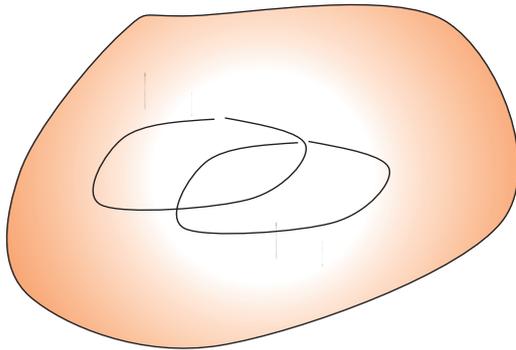
$$p(\mathbf{x}) \in \mathbb{P} \quad \int_{\mathcal{D}} p(\mathbf{x}) d\mathbf{x} = 1 \quad p(\mathbf{x}) \geq 0$$



$$\text{Probability} = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$



# Probability and Probability Density



$$\begin{aligned} P(\mathbf{x} \in \mathcal{R}_1 \cup \mathbf{x} \in \mathcal{R}_2) &= \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x}) d\mathbf{x} \\ &= P(\mathcal{R}_1) + P(\mathcal{R}_2) - P(\mathcal{R}_1 \cap \mathcal{R}_2) \end{aligned}$$

Event is defined infinitesimally:

$\mathcal{R}$ : set of infinitesimal events



## Can you explain the meaning of these functions?

$$P(X = 1)$$

$$P(X = 1|Y = 2)$$

$$p(x = 1) \quad \text{Compare with } P(x = 1)?$$

$$p(x = 1|y = 2)$$



# Bayes Optimal Classifier

- Our ultimate goal is *not a zero error*.

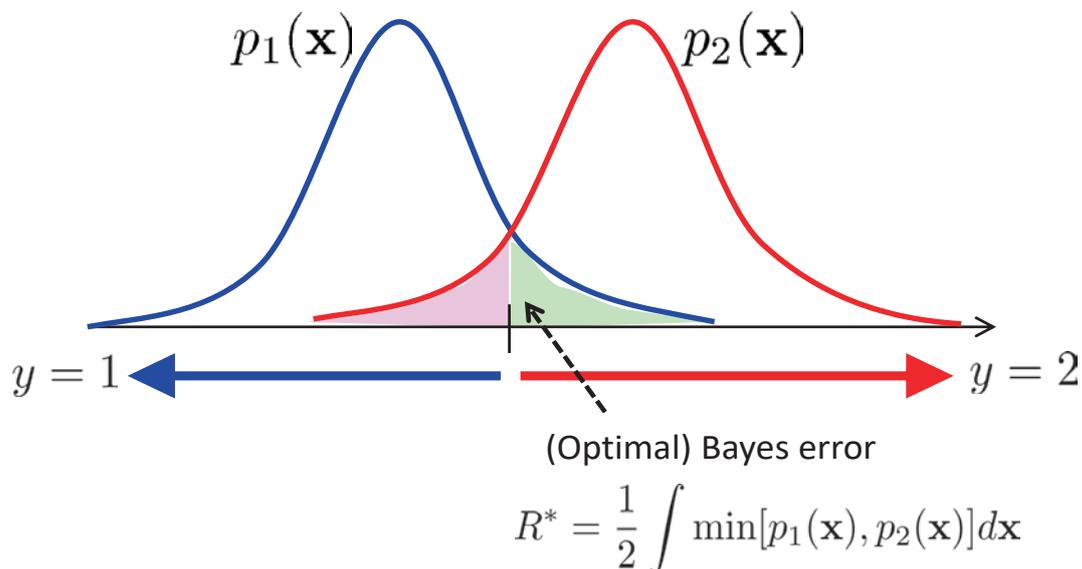
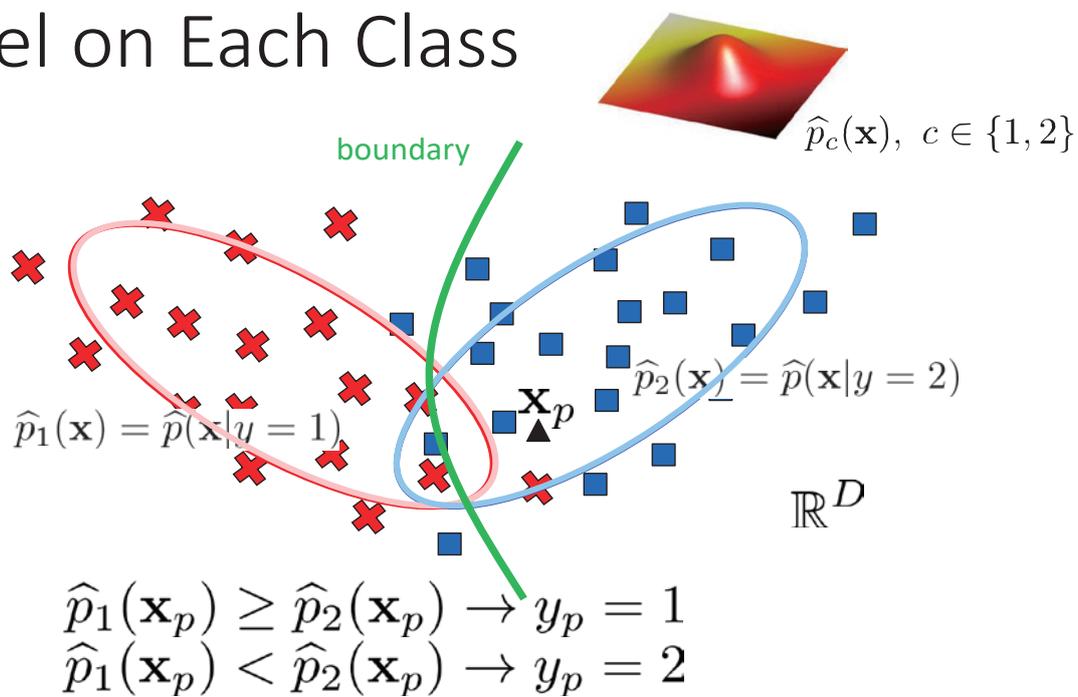


Figure credit: Masashi Sugiyama



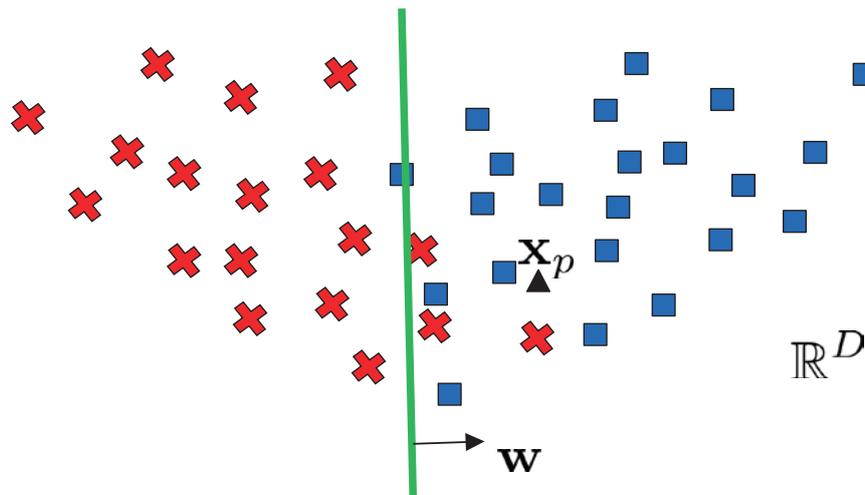
# Model on Each Class



- Model: Assumptions on the Class-conditional density



# Model on Discriminative Function



- Model: Assumptions on the boundary and optimize the boundary directly from data

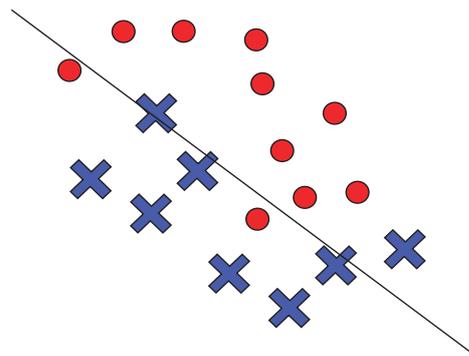


# Consistent but Asymptotically Nonplausible

- Discriminative model

$$P(Y|\mathbf{x})$$

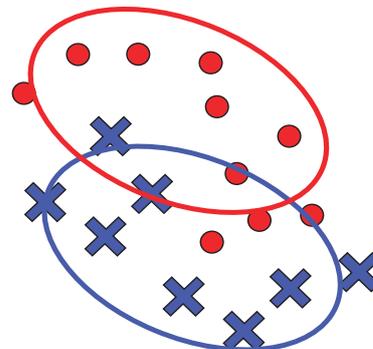
Linear discriminant models (logistic regression, ...)



- Generative model

$$P(\mathbf{x}|Y)$$

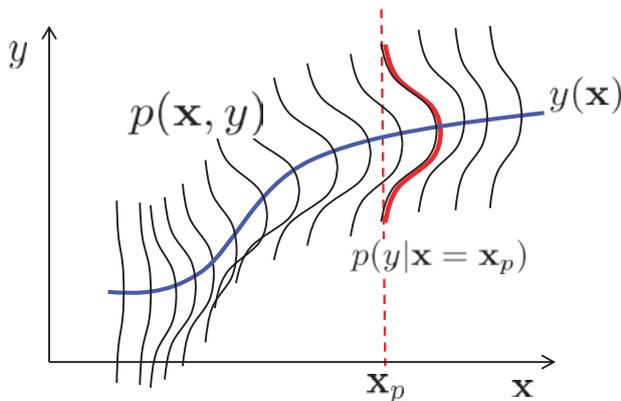
Gaussian models,  
Fisher Discriminant Analysis  
Naïve-Bayes models,  
Graphical models,  
...



# Optimal Regression

- Minimizing mean square error

$$y(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = \int y p(y|\mathbf{x}) dy$$



Minimize

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

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

→ Minimized when  $y(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$

# Model for Regression

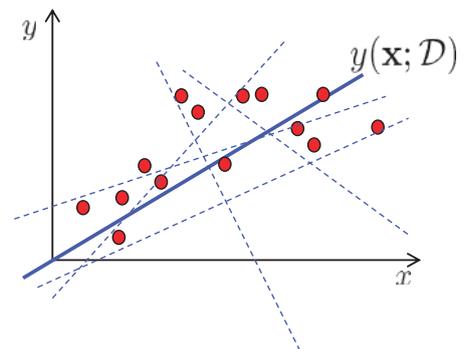
- Obtain regression function from data  $y(\mathbf{x}; \mathcal{D}) \in \mathcal{H}$

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

- Choose a model  $\mathcal{H}$  where the following expectation is minimized:

$$\mathbb{E}_{\mathcal{D}} \left[ \overset{\text{function with minimum } \hat{L}}{\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2} \right]$$

- Minimized for  $y(\mathbf{x}; \mathcal{D}) = \mathbb{E}[y|\mathbf{x}]$



- Bias-Variance tradeoff

$$\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2 = \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - \mathbb{E}[y|\mathbf{x}]\}^2$$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}} \left[ \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\}^2 \right] &= \mathbb{E}_{\mathcal{D}} \left[ \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 \right] + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - \mathbb{E}[y|\mathbf{x}]\}^2 \\ &\quad \nearrow \text{Variance} \qquad \qquad \qquad \nearrow \text{Bias}^2 \end{aligned}$$

# Several Rules

$$\sum_{X_i \in \text{all disjoint set}} P(X = X_i) = 1$$

$$\sum_{X_i \in \text{all disjoint set}} P(X = X_i | Z = Z_j) = 1$$

$$\sum_{Z_j \in \text{all disjoint set}} P(X = X_i | Z = Z_j) = ?$$



# For More Than Two Random Variables

- For three disjoint sets  $X_1, X_2, X_3$  for a random variable  $X$  and another three disjoint sets  $Y_1, Y_2, Y_3$  for a random variable  $Y$ :

$Y \backslash X$	$X_1$	$X_2$	$X_3$	
$Y_1$	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
$Y_2$	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
$Y_3$	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	1



# Conditional Probability

$Y \backslash X$	$X_1$	$X_2$	$X_3$	
$Y_1$	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
$Y_2$	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
$Y_3$	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	<b>1</b>

$$P(X = X_1 | Y = Y_1) = \frac{P(X_1, Y_1)}{P(X_1, Y_1) + P(X_2, Y_1) + P(X_3, Y_1)}$$

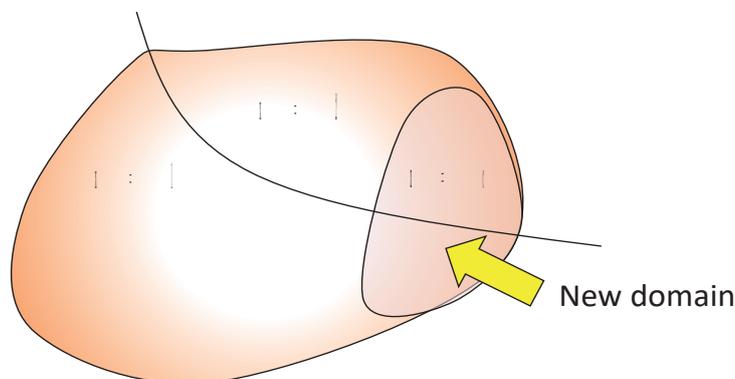
$$= \frac{P(X_1, Y_1)}{P(Y_1)}$$



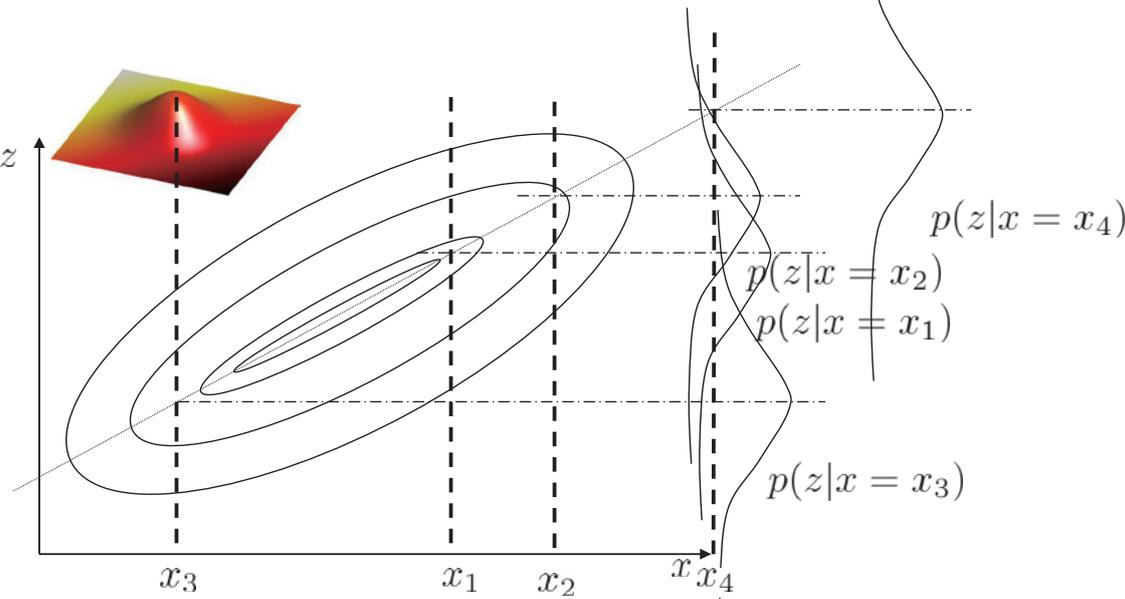
# Conditional Probability Density

$$p(\mathbf{x}, \mathbf{z}) \quad \mathbf{x} \in \mathbb{R}^{D_x}, \mathbf{z} \in \mathbb{R}^{D_z}$$

$$\rightarrow p(\mathbf{x} | \mathbf{z} = c) = \frac{p(\mathbf{x}, \mathbf{z} = c)}{\int p(\mathbf{x}, \mathbf{z} = c) d\mathbf{x}}$$

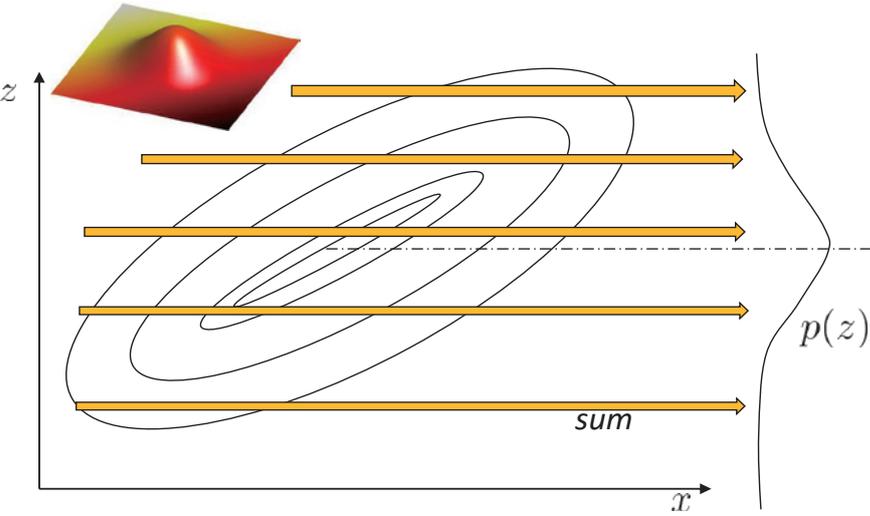


# Conditional Probability Density



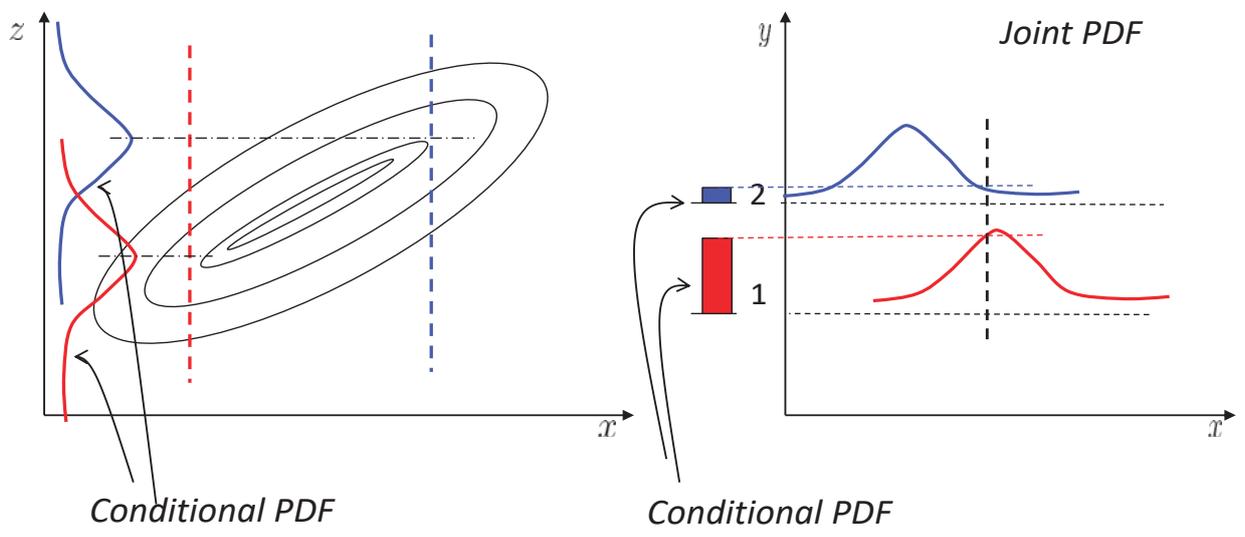
$$p(z|x = x_0) = \frac{p(z, x = x_0)}{\int_{x=x_0} p(z, x) dz} = \frac{p(z, x)}{p(x)}$$

# Marginal Probability Density

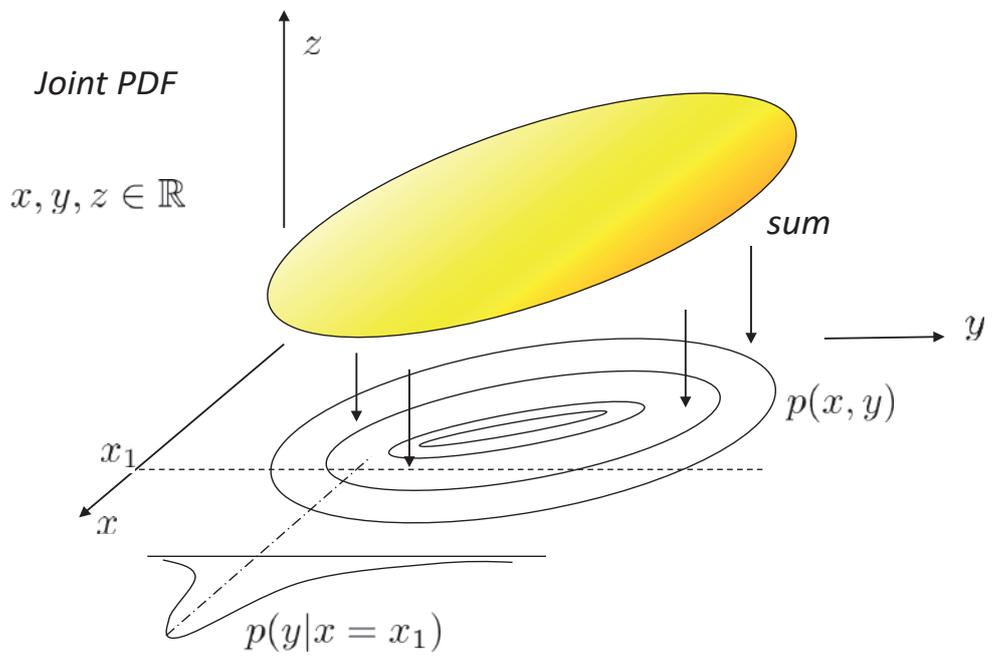


$$p(z) = \int p(z, x) dx$$

# Marginal Probability Density and Conditional Probability Density in Machine Learning

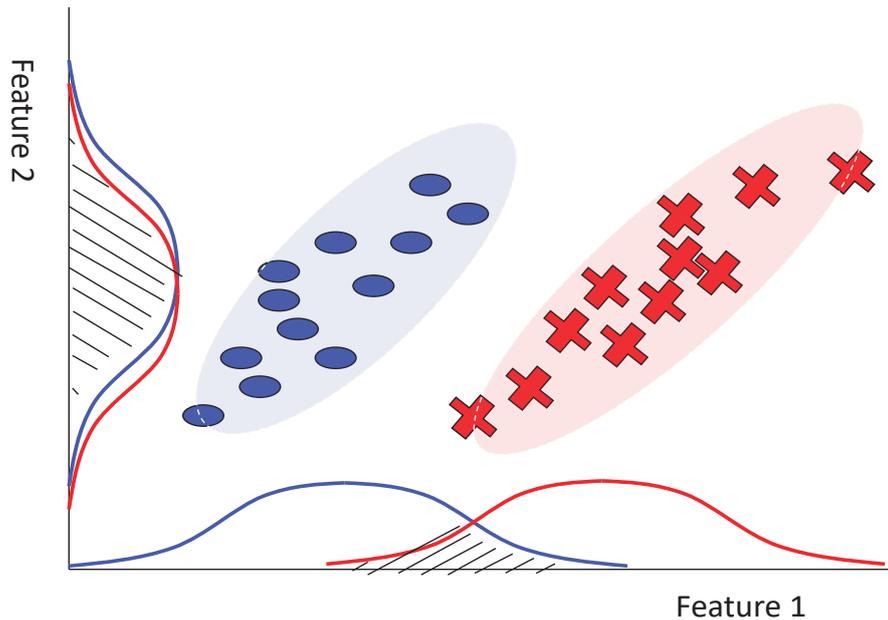


# Marginal Probability Density and Conditional Probability Density in Machine Learning



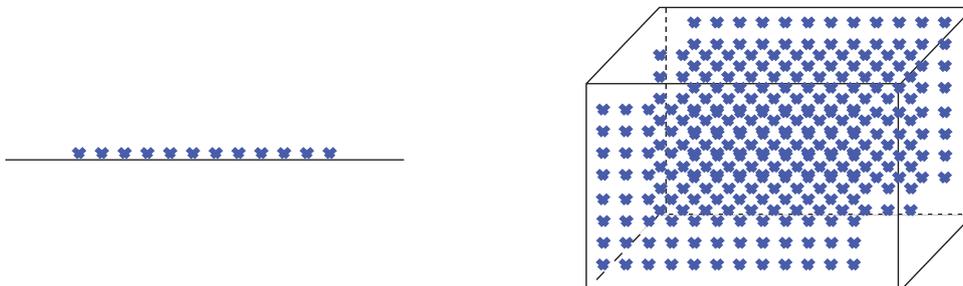
# Benefits of Using High Dimensionalities

- Feature 1 and Feature 2 have correlation



# Curse of Dimensionality

- To achieve the 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.

# Gaussian Density Function



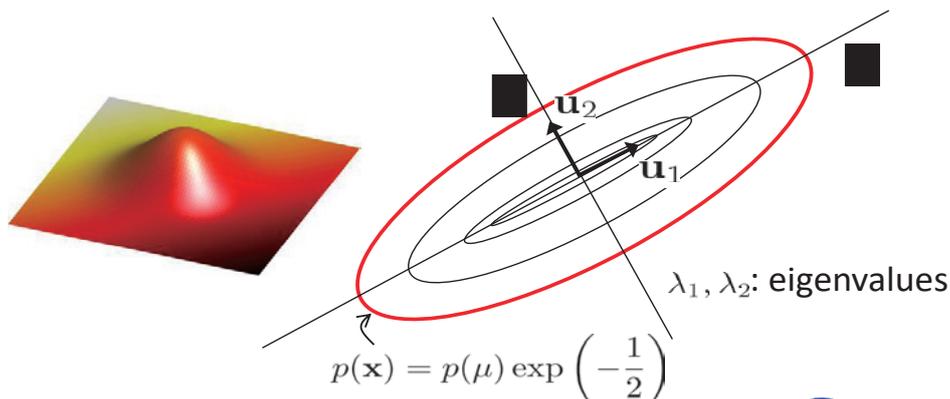
# Gaussian Random Variable

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

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

Principal axes are the eigenvector directions of  $\Sigma$

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



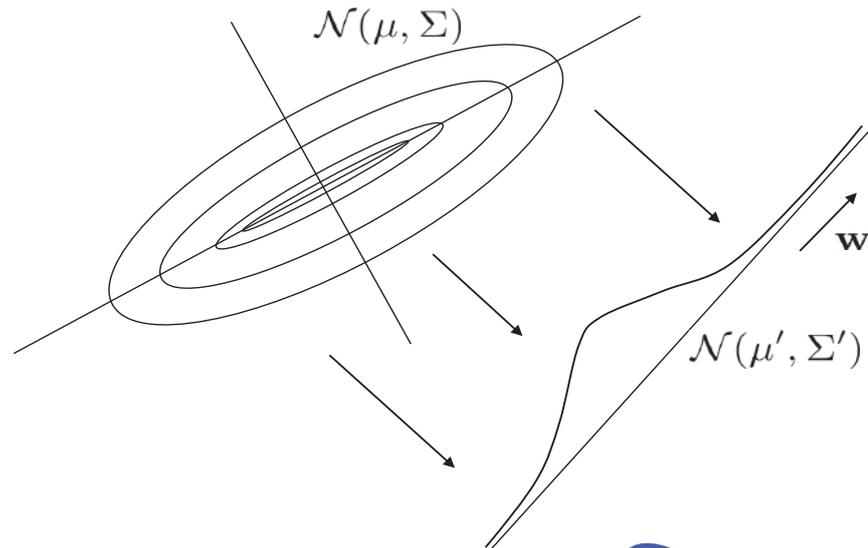
# Gaussian Random Variable - Projection

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

Projection to any direction is Gaussian.

$$\mu' = \mathbf{w}^\top \mu$$

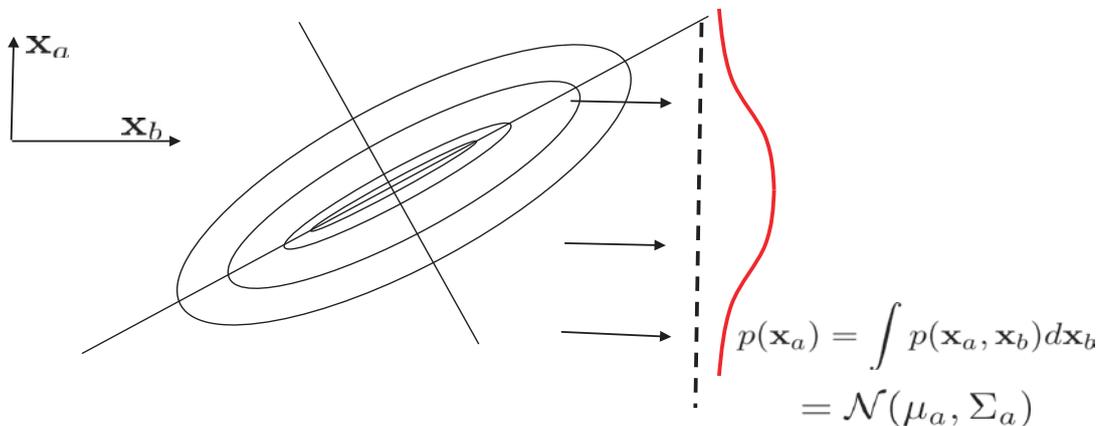
$$\Sigma' = \mathbf{w}^\top \Sigma \mathbf{w}$$



# Gaussian Random Variable – Marginal

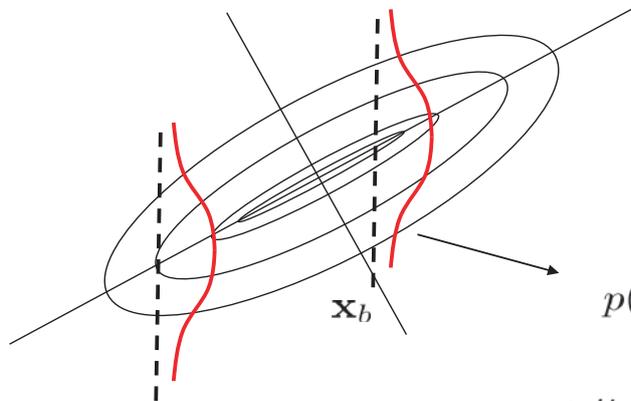
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

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



# Gaussian Random Variable – Conditional

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



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

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

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

# Gaussian Parameter Estimation and Inference - Simple Example

- $\mathbf{x} \in \mathbb{R}^D$  and  $y \in \mathbb{R}$  are jointly Gaussian.
- Using  $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ , estimate

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_y \\ \hat{\mu}_{\mathbf{x}_a} \\ \hat{\mu}_{\mathbf{x}_b} \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_y & \hat{\Sigma}_{y\mathbf{x}_a} & \hat{\Sigma}_{y\mathbf{x}_b} \\ \hat{\Sigma}_{\mathbf{x}_a y} & \hat{\Sigma}_{\mathbf{x}_a} & \hat{\Sigma}_{\mathbf{x}_a \mathbf{x}_b} \\ \hat{\Sigma}_{\mathbf{x}_b y} & \hat{\Sigma}_{\mathbf{x}_b \mathbf{x}_a} & \hat{\Sigma}_{\mathbf{x}_b} \end{pmatrix}$$

where

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}.$$

- For a new datum  $\mathbf{x}$  with missing  $\mathbf{x}_b$ ,

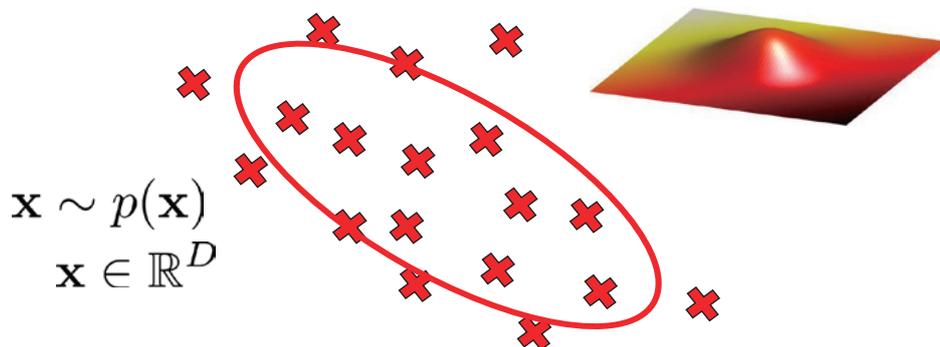
$$\hat{p}(y | \mathbf{x}_a) = \mathcal{N}\left(\hat{\mu}_y + \hat{\Sigma}_{y\mathbf{x}_a} \hat{\Sigma}_{\mathbf{x}_a}^{-1} (\mathbf{x}_a - \hat{\mu}_a), \hat{\Sigma}_y - \hat{\Sigma}_{y\mathbf{x}_a} \hat{\Sigma}_{\mathbf{x}_a}^{-1} \hat{\Sigma}_{\mathbf{x}_a y}\right)$$

# Parameter Estimation



## Motivation – Parameter Estimation

- Parameter estimation is an optimization problem

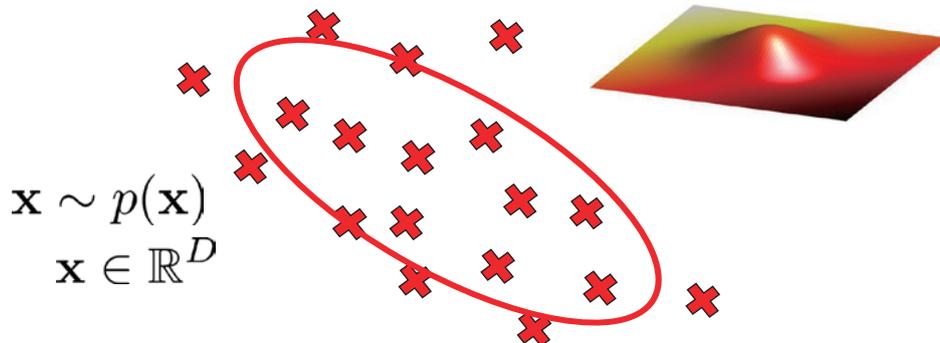


$\hat{p}(\mathbf{x})$ : estimated probability density function,  
in other words, density function that fits data the most



# Maximum Likelihood Estimation

- Parameter estimation is an optimization problem



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

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



# Maximum Likelihood for Gaussian

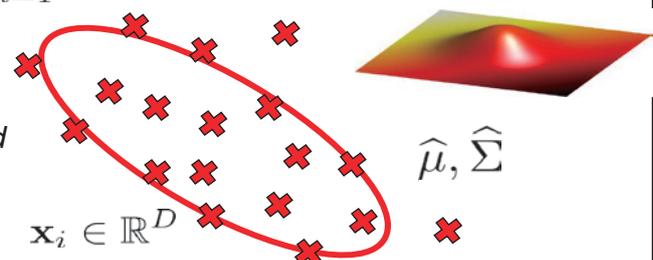
$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

- With optimal parameters satisfying

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

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

*Empirical mean and empirical covariance are the maximum likelihood solutions.*



# Maximum Likelihood for Gaussian

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

$$\nabla_{\theta} \ln p(X|\theta) = \vec{0}$$

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

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

# Covariance Estimation

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

- In high-dimensional space



*(D + 1)D/2 number of parameters for covariances*

## Maximum A Posteriori (MAP) Estimation

- MAP estimation

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

- 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

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

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

- 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

$$\begin{aligned} \prod_{i=1}^N p(x_i|\mu)p(\mu) &= \left[ \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \right] \\ &\quad \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\sum \frac{(x_i - \mu)^2}{\sigma^2} + \frac{\mu - \mu_0}{\sigma_0^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mu^2 \left[\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right] - 2\mu \left[\frac{1}{\sigma^2} \sum x_i + \frac{\mu_0}{\sigma_0}\right]\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_n^2}(\mu - \mu_n)^2\right) \end{aligned}$$



## Maximum A Posteriori (MAP) Estimation for Gaussian

- Posterior density

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

- Caution: Posterior of  $\mu$ , not the density function of  $\mathcal{X}$

- MAP of  $\mu$  = Mean of  $\mu$  =  $\mu_n$

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



# MLE vs. MAP

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

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

$$\mu_n = \frac{25}{5 \cdot 5 + 3} \hat{\mu}_{ML} + \frac{3}{5 \cdot 5 + 3} \mu_0$$

Dominant

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



# MLE vs. MAP

- For Gaussian
  - When we have a few outliers

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

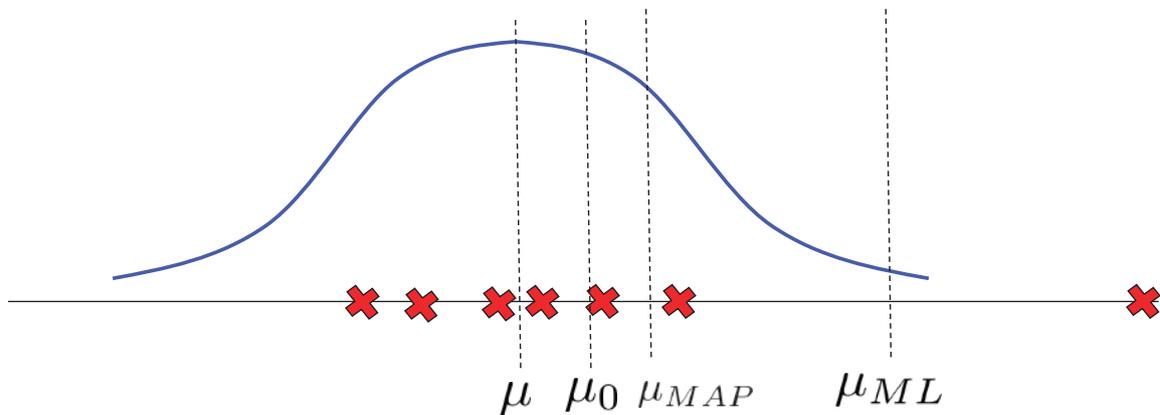
$$\mu_n = \frac{25}{5 \cdot 5 + 100} \hat{\mu}_{ML} + \frac{100}{5 \cdot 5 + 100} \mu_0$$

Dominant (learn from )  $\mu_0$

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



# MLE vs. MAP



# Bayesian Integration

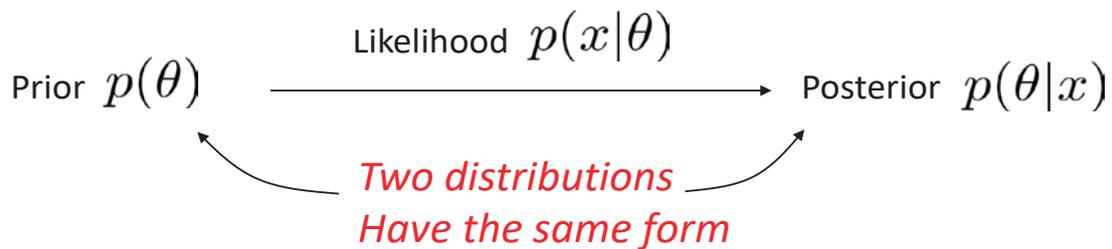
- The final standard method of prediction is to use Bayesian inference instead of estimating the parameter point.
  - Do not insert the point estimate  $\hat{\mu}_{MAP}$  directly, but marginalize.

$$\begin{aligned}
 p(x|X) &= \int p(x|\mu)p(\mu|X)d\mu \\
 &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right) d\mu \\
 &= \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_n^2)}} \exp\left(-\frac{1}{2(\sigma^2 + \sigma_n^2)}(x-\mu)^2\right) \\
 &= \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2) \quad \text{Uncertainty of } \mu
 \end{aligned}$$

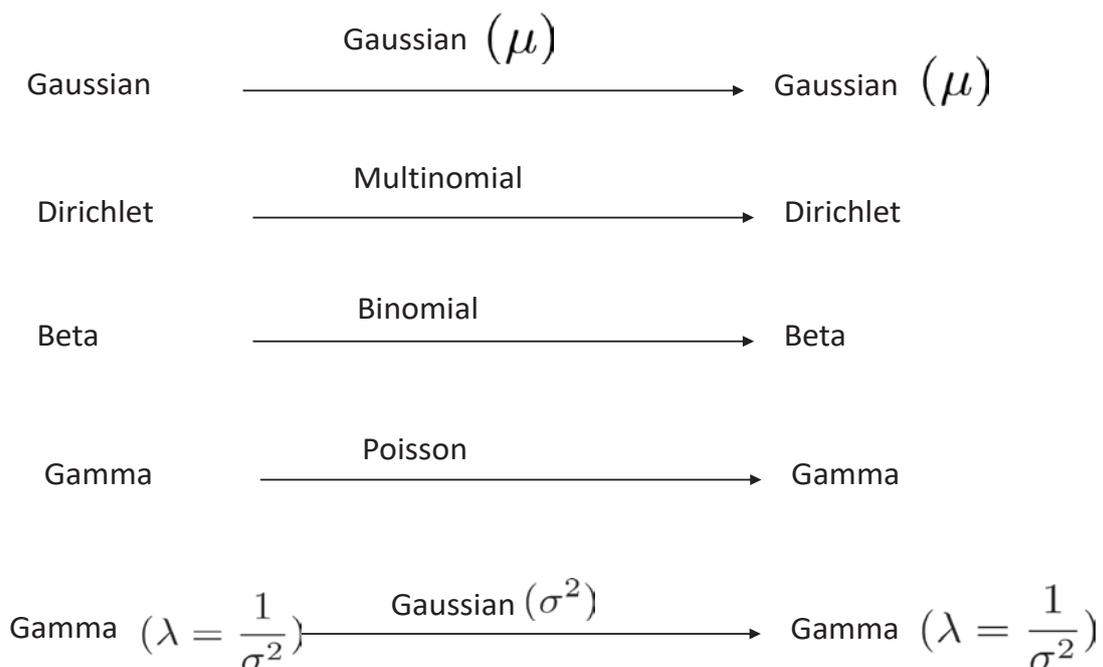


# Conjugate Priors

- Given a likelihood pdf,  $p(x|\theta)$  posterior  $p(\theta|x)$  has the same form as the prior  $p(\theta)$ .



# Conjugate Priors



# Kullback-Leibler Divergence

$$KL(p_e || p_\theta) = - \int p_e \log \frac{p_\theta}{p_e} d\mathbf{x} \quad \begin{array}{l} p_e: \text{Empirical density function} \\ p_\theta: \text{Model density function} \end{array}$$

$$= - \int [p_e \log p_\theta - p_e \log p_e] d\mathbf{x}$$

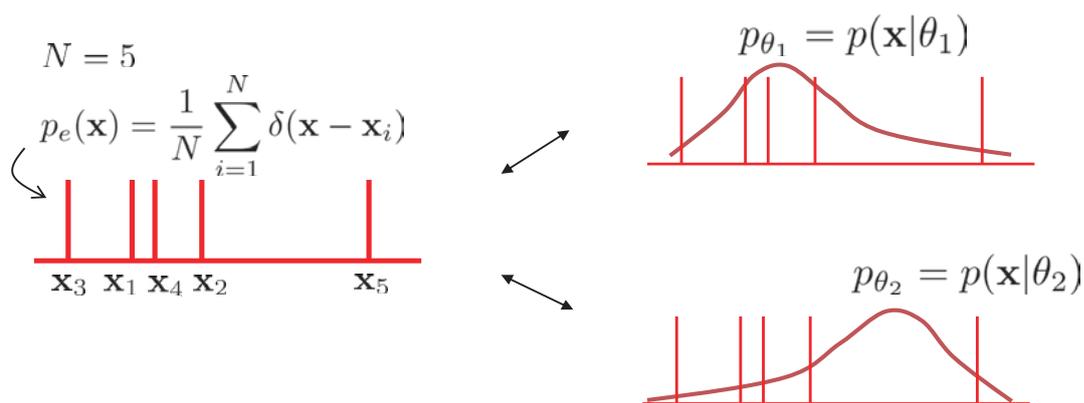
$$p_e = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)$$

$$\arg \min_{p_\theta} KL(p_e || p_\theta) = \arg \min_{p_\theta} - \int \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i) \log p_\theta(\mathbf{x}) d\mathbf{x}$$

$$= \arg \max_{p_\theta} \frac{1}{N} \sum_{i=1}^N \log p_\theta(\mathbf{x}_i)$$

$$= \arg \max_{p_\theta} \log \prod_{i=1}^N p_\theta(\mathbf{x}_i) = \arg \max_{p_\theta} p(\mathcal{D} | \theta)$$

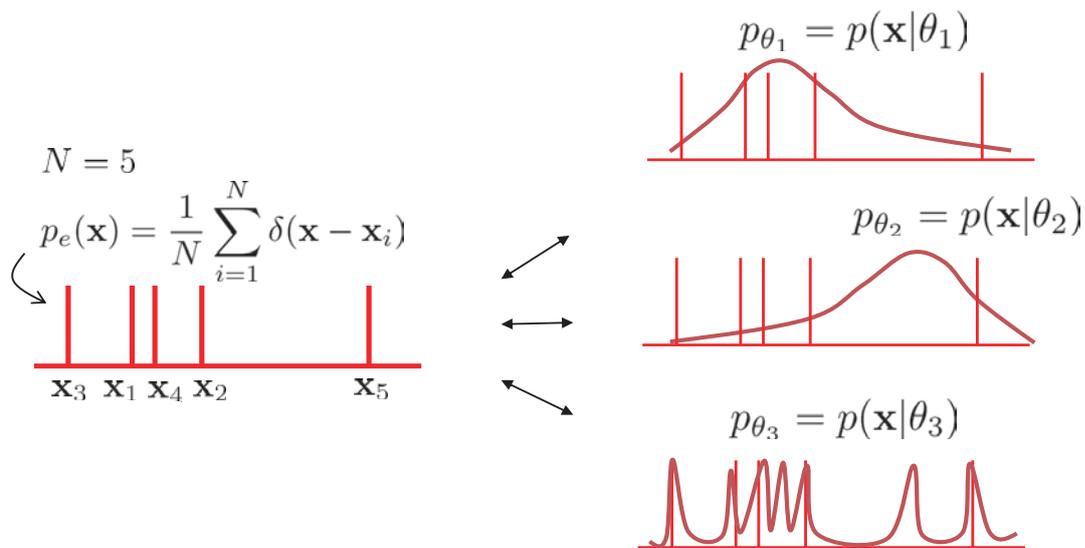
# Kullback-Leibler Divergence



KL Divergence:  $KL(p_e || p_{\theta_1}) < KL(p_e || p_{\theta_2})$

Likelihood:  $p(\mathcal{D} | \theta_1) > p(\mathcal{D} | \theta_2)$

# Kullback-Leibler Divergence



$$\theta_3 = \arg \max_{\theta} p(\mathcal{D}|\theta)$$

Model with complex function will capture the noise.



## Thank you

Yung-Kyun Noh

nohyung@hanyang.ac.kr

