KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

An Introduction to Probabilistic Modeling

노영균 _ 한양대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 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 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의개요

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	Asssistant Professor
Affiliation	Hanyang University

▶ **Contact Information**

Email	nohyung@hanyang.ac.kr
Phone Number	02-2220-1409

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

SBi 한국생명정보학회 Korean Society for Bioinformatics

AAIS

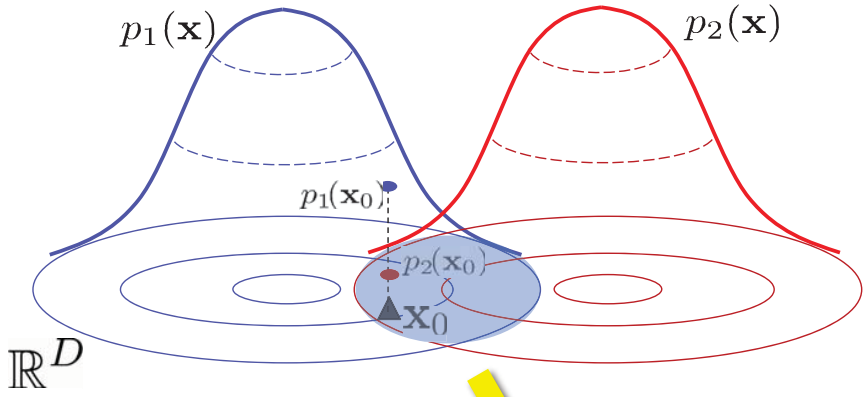
http://aais.hanyang.ac.kr

Contents

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

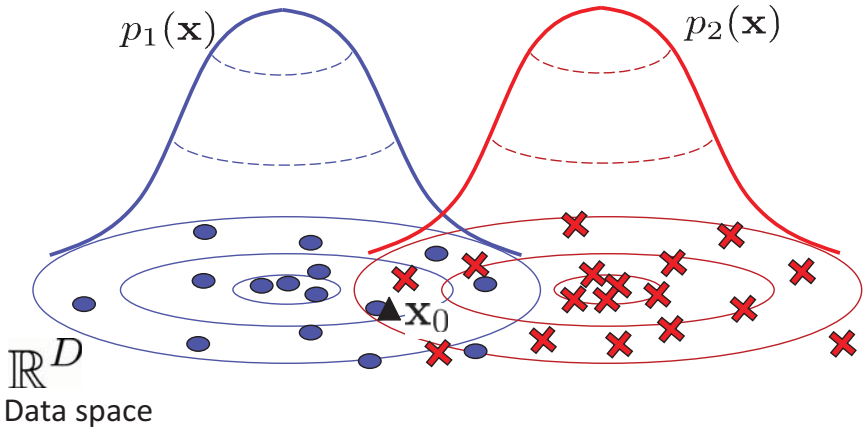
Probabilistic Assumption and Bayes Classification

- Bayes Error



$$E_{Bayes} = \frac{1}{2}\int \min[p_1, p_2] d\mathbf{x}$$

Probabilistic Assumption and Bayes Classification



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

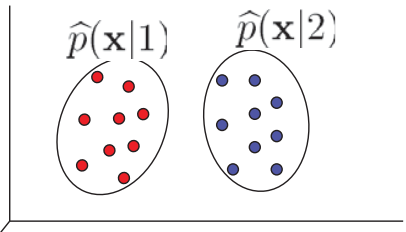
Generative vs. Discriminative (1/2)

- Generative Learning
  - ➢ Interested in joint probability

$$\widehat{p}(\mathbf{x}, y)$$

$$\widehat{p}(y) \qquad \widehat{p}(\mathbf{x}|y)$$

→ Apply class conditional model



- Prediction
  - ➢ Bayes Rule

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

Generative vs. ***Discriminative*** (2/2)

- Discriminative learning
  - ➢ "*NOT*" interested in joint probability

- Conditional probability learning

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

- Learn prediction function by minimizing the empirical loss function

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

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

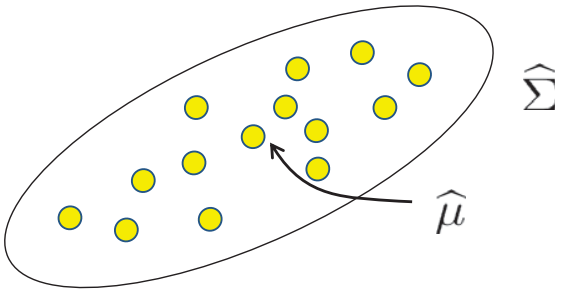
Learn Density Function from Data

- 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)

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

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

$$\widehat{p}_1(\mathbf{x}) \gtrless \widehat{p}_2(\mathbf{x}) \Leftrightarrow \frac{\widehat{p}_2(\mathbf{x})}{\widehat{p}_1(\mathbf{x})} \gtrless 1$$

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

For Two Gaussian Data (2/2)

- With a Homoscedastic Assumption

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

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

The problem reduces to

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

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

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

→ Fisher Discriminant Analysis

In Terms of the Posterior

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

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

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)

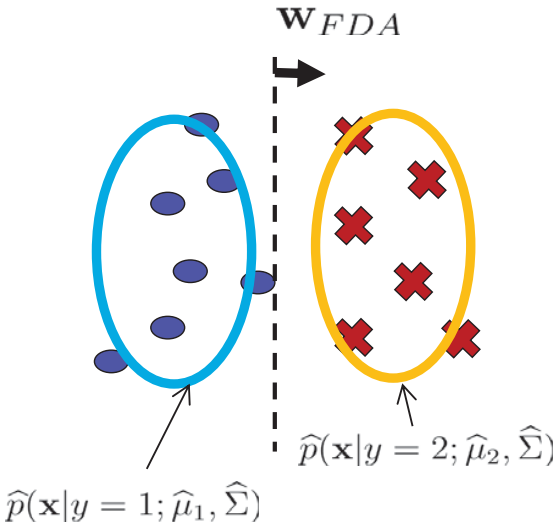
$$\mathbf{w}^T\mathbf{x} - b \gtrless 0$$

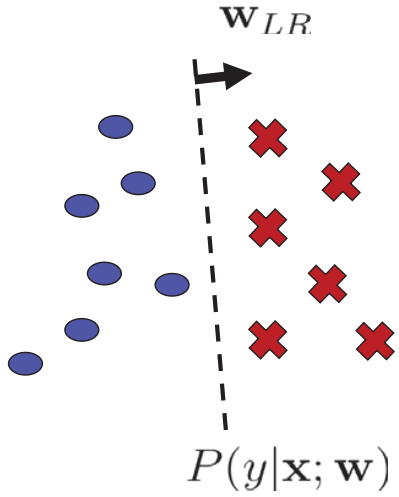
  - FDA solution: Bayes classifier with class-conditional model
  - Logistic regression: Discriminative adaptation of a discriminative function

  - Question: *Are the results the same or not?*

FDA and Logistic Regression

- *Are the results the same or not?*



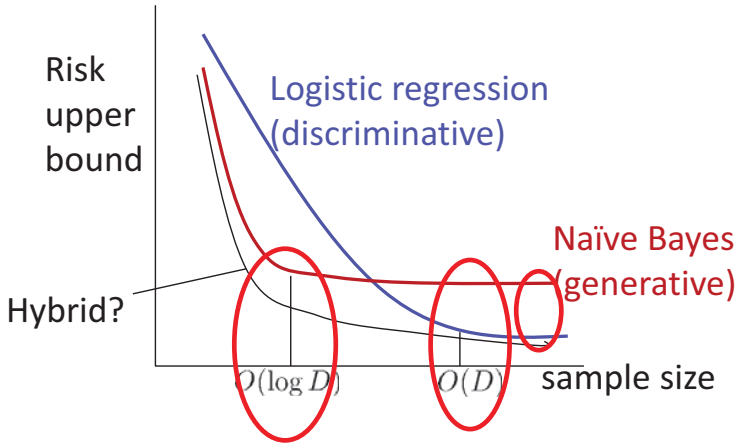


$$\mathbf{w}_{FDA} = \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2)$$

$$\mathbf{w}_{LR} = \arg\max_{\mathbf{w}} P(y|\mathbf{x};\mathbf{w})$$

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)

$$p(\mathbf{x}_2|\mathbf{x}_1) \qquad P(y|\mathbf{x})$$

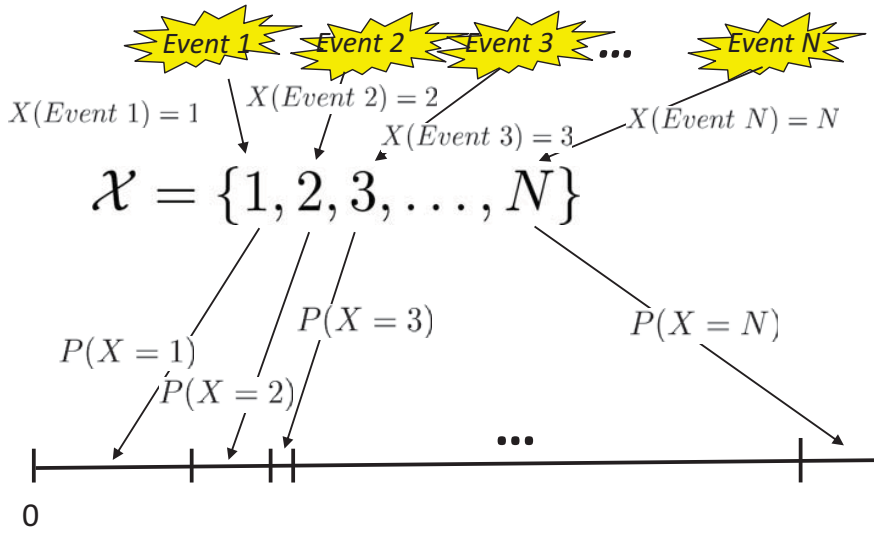
$$\mathbf{x}_1 \in \mathbb{R}^{D_1}, \mathbf{x}_2 \in \mathbb{R}^{D_2}, \mathbf{x} \in \mathbb{R}^{D}, y \in \{1, 2\}$$

- 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

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

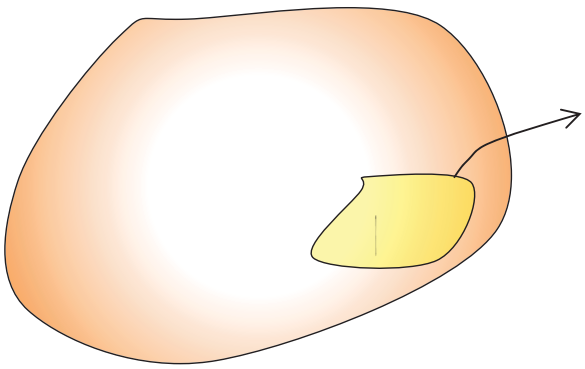
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

$$X_1 = \{1, 2, 3, 4\}, \; X_2 = \{3, 4, 5\}$$
$$P(1, 2, 3, 4, 5) = P(1, 2, 3, 4) + P(3, 4, 5) - P(3, 4)$$

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

Probability and Probability Density

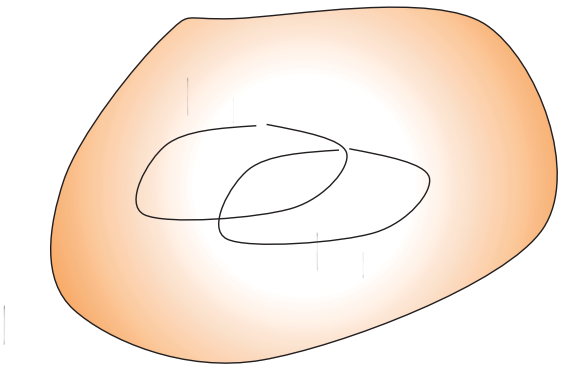
$$p(\mathbf{x}) \in \mathbb{P} \qquad \int_{\mathcal{D}} p(\mathbf{x}) d\mathbf{x} = 1 \qquad p(\mathbf{x}) \geq 0$$



$$P(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$

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

Probability and Probability Density



$$P(\mathbf{x} \in \mathcal{R}_1 \cup \mathbf{x} \in \mathcal{R}_2) = \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x}) d\mathbf{x}$$

$$= P(\mathcal{R}_1) + P(\mathcal{R}_2) - P(\mathcal{R}_1 \cap \mathcal{R}_2)$$

Event is defined infinitesimally:

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

Can you explain the meaning of these functions?

$P(X = 1)$

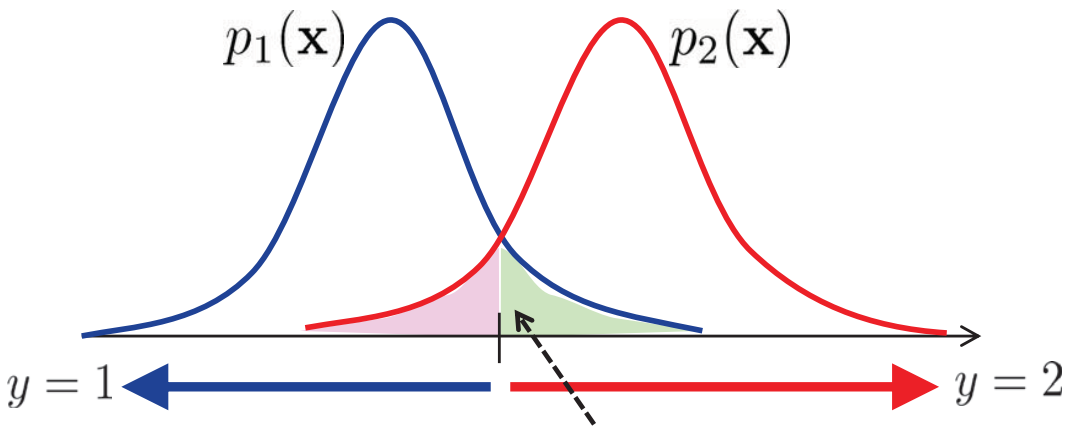
$P(X = 1|Y = 2)$

$p(x = 1)$ Compare with $P(x = 1)$?

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

Bayes Optimal Classifier

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

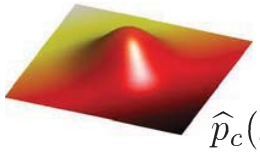


(Optimal) Bayes error

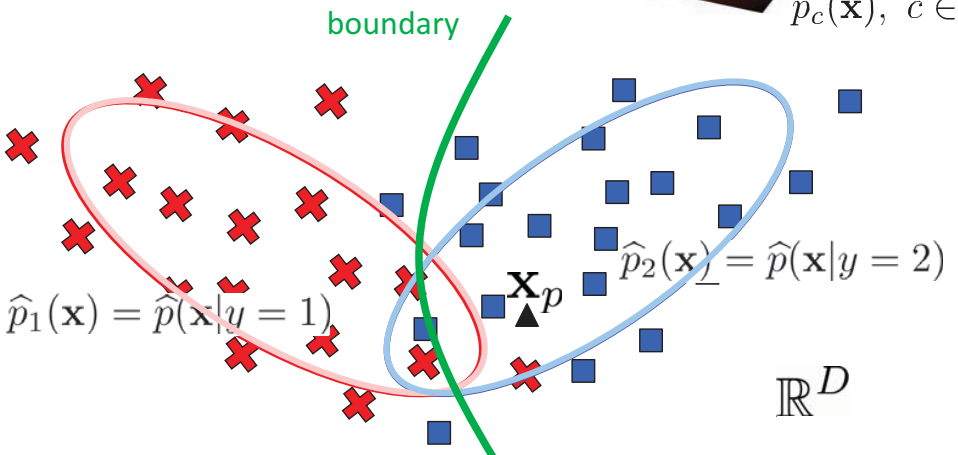
$$R^* = \frac{1}{2}\int \min[p_1(\mathbf{x}), p_2(\mathbf{x})]d\mathbf{x}$$

Figure credit: Masashi Sugiyama

Model on Each Class



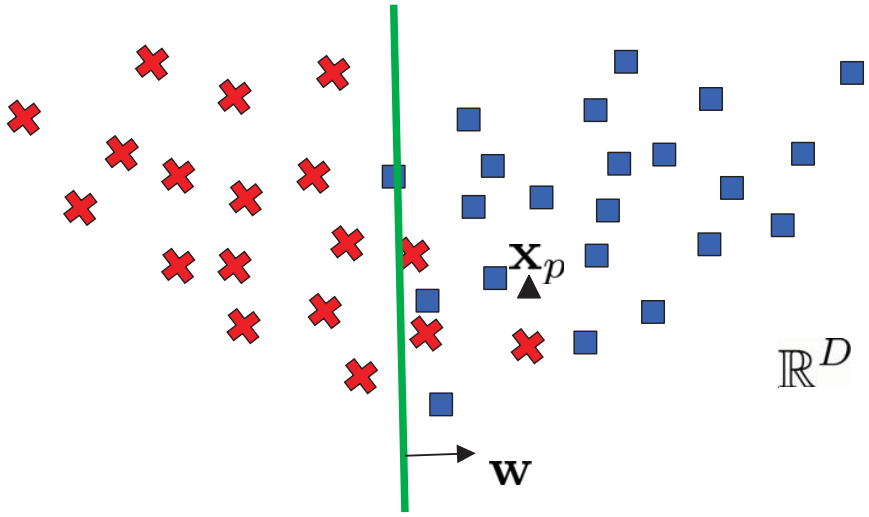
$$\widehat{p}_c(\mathbf{x}),\ c \in \{1, 2\}$$



$$\widehat{p}_1(\mathbf{x}_p) \geq \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 1$$
$$\widehat{p}_1(\mathbf{x}_p) < \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 2$$

- Model: Assumptions on the Class-conditional density

Model on Discriminative Function



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

Consistent but Asymptotically Nonplausible

- Discriminative model

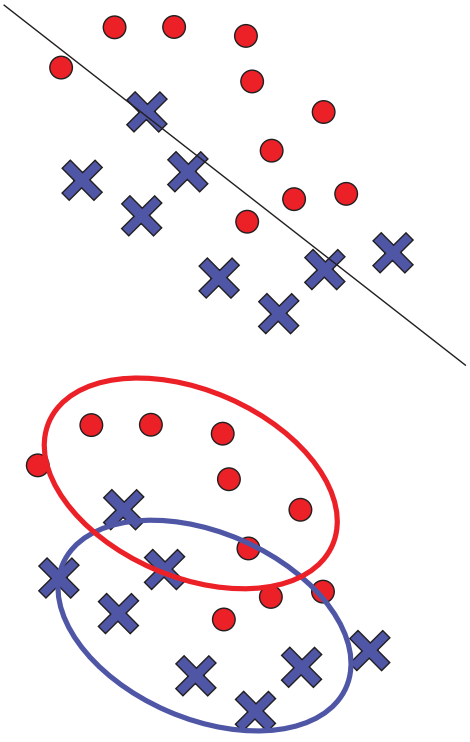
$$P(Y|\mathbf{x})$$

Linear discriminant models (logistic regression, …)

- Generative model

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

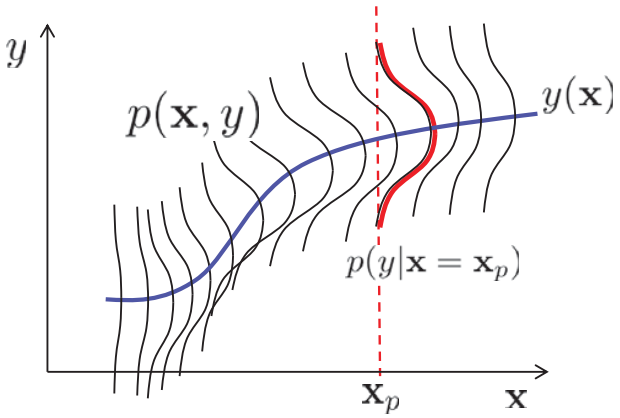
Gaussian models,
Fisher Discriminant Analysis
Naïve-Bayes models,
Graphical models,
…



Optimal Regression

- Minimizing mean square error

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



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 $y(\mathbf{x}) = \mathbb{E}\left[y|\mathbf{x}\right]$

Model for Regression

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

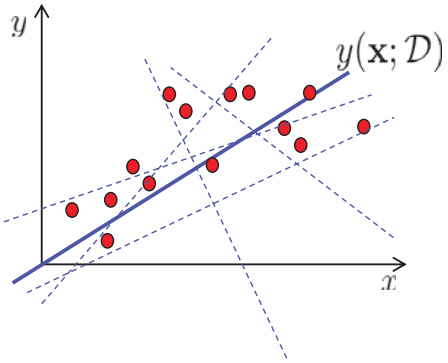
$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \sim p(\mathbf{x}, y)$$

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

function with minimum $\widehat{L}$

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

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



- Bias-Variance tradeoff

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

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

Variance Bias^2

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

Several Rules

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

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

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

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 \ X	X_1	X_2	X_3	$P(X \in \{X_1, X_2\}, Y \in Y_1)$
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

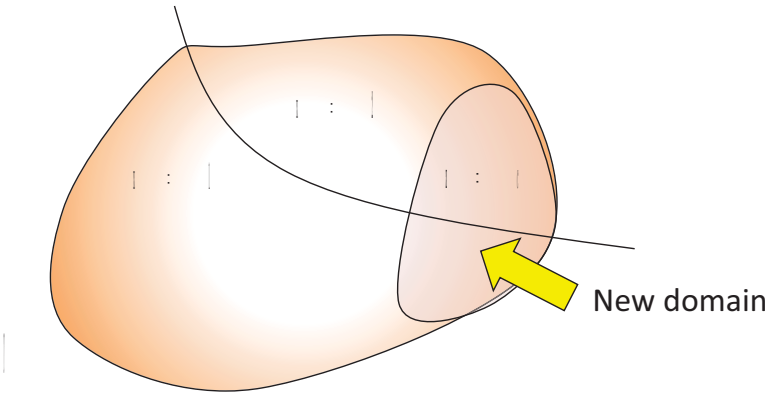
Y \ X	X_1	X_2	X_3	
Y_1	$P(X_1, Y_1)$	$P(X_2, Y_1)$	$P(X_3, Y_1)$	$P(Y_1)$
Y_2	$P(X_1, Y_2)$	$P(X_2, Y_2)$	$P(X_3, Y_2)$	$P(Y_2)$
Y_3	$P(X_1, Y_3)$	$P(X_2, Y_3)$	$P(X_3, Y_3)$	$P(Y_3)$
	$P(X_1)$	$P(X_2)$	$P(X_3)$	1

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

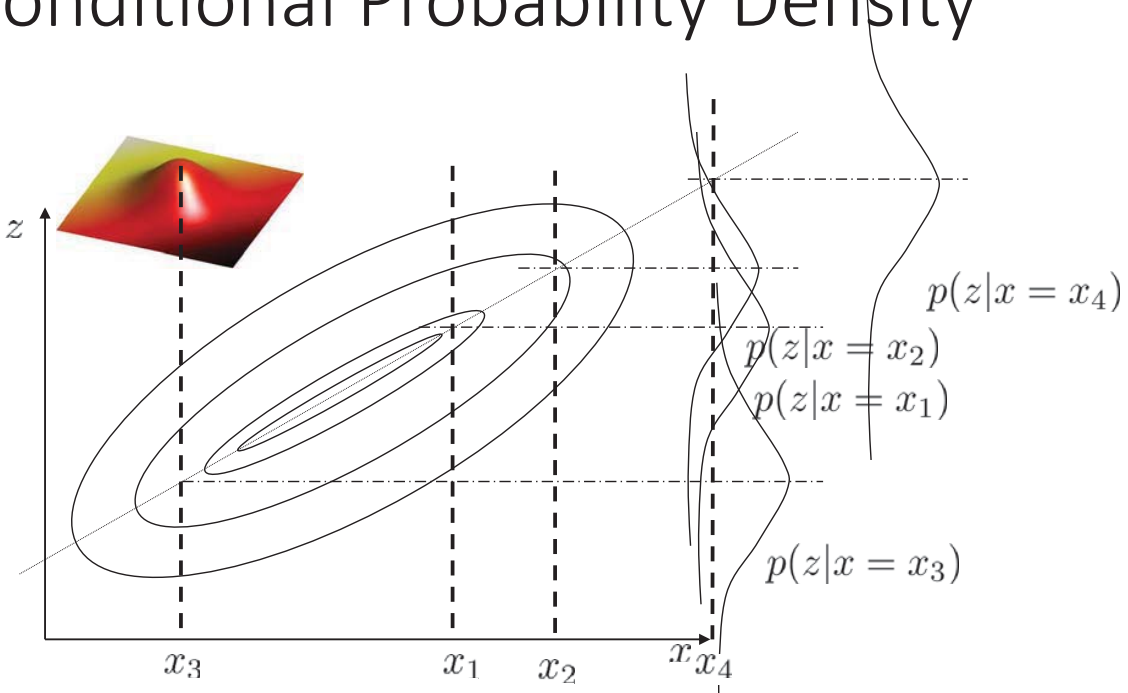
Conditional Probability Density

$$p(\mathbf{x}, \mathbf{z}) \qquad \mathbf{x} \in \mathbb{R}^{D_\mathbf{x}}, \mathbf{z} \in \mathbb{R}^{D_\mathbf{z}}$$

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

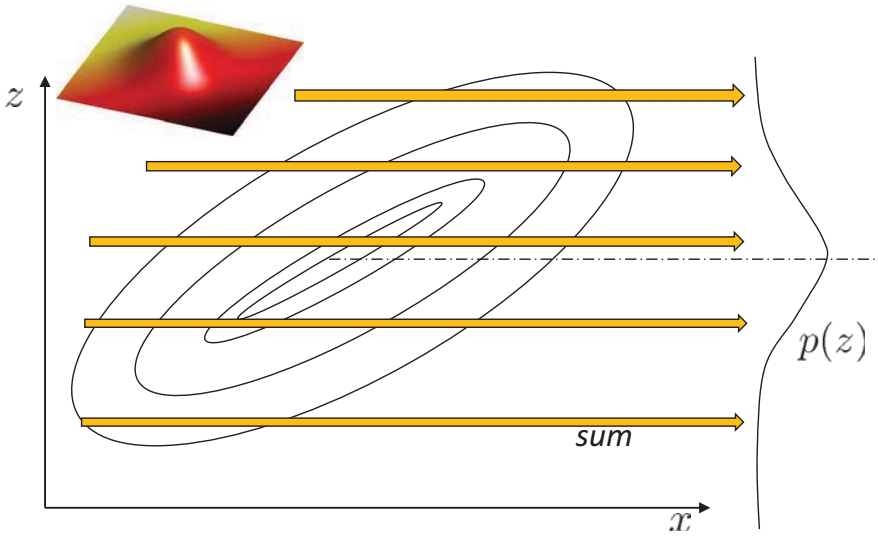


Conditional Probability Density



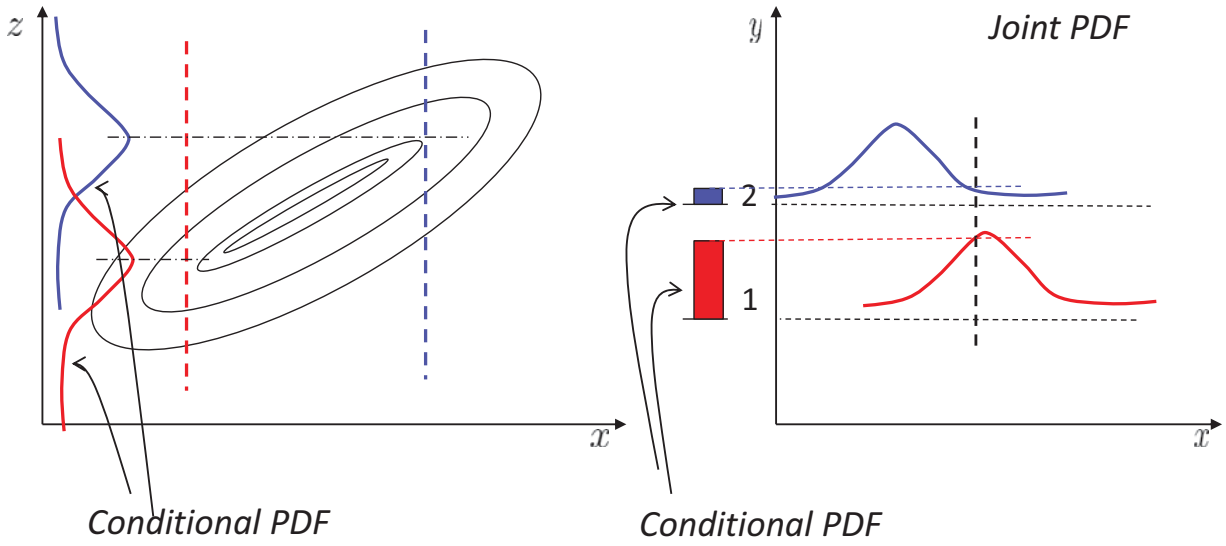
$$p(z|x = x_0) = \frac{p(z, x = x_0)}{\int_{x=x_0} p(z, x)dz} = \frac{p(z, x)}{p(x)}$$

Marginal Probability Density

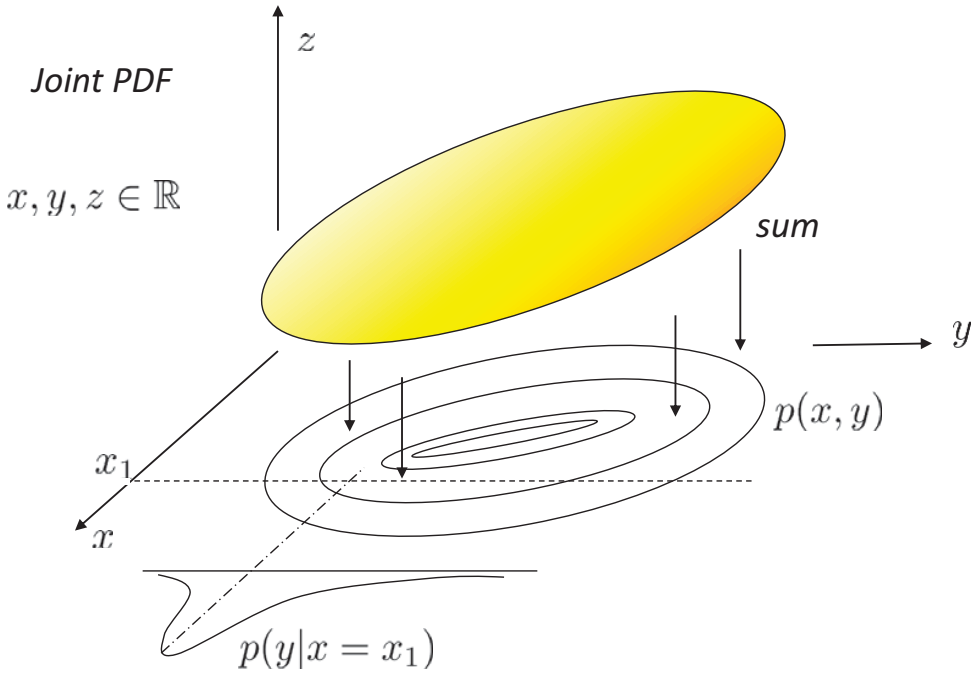


$$p(z) = \int p(z, x)dx$$

Marginal Probability Density and Conditional Probability Density in Machine Learning

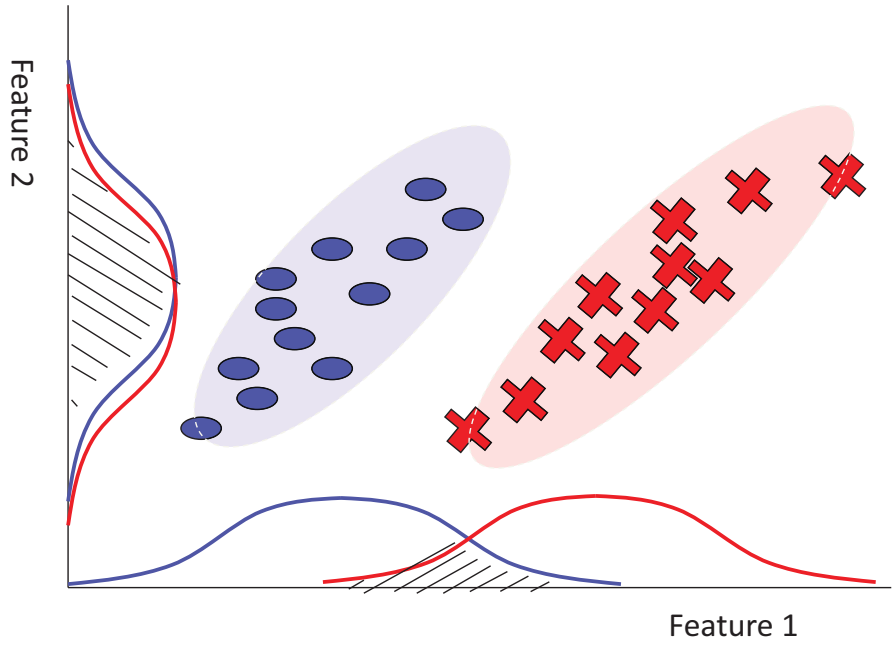


Marginal Probability Density and Conditional Probability Density in Machine Learning



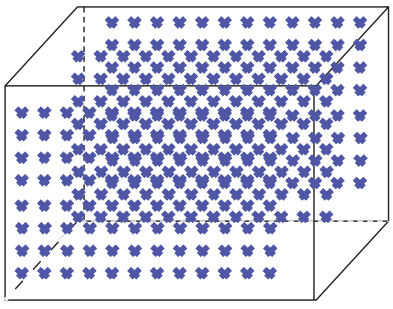
Benefits of Using High Dimensionalities

- 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

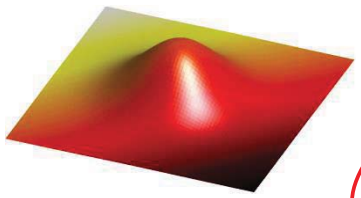
Gaussian Random Variable

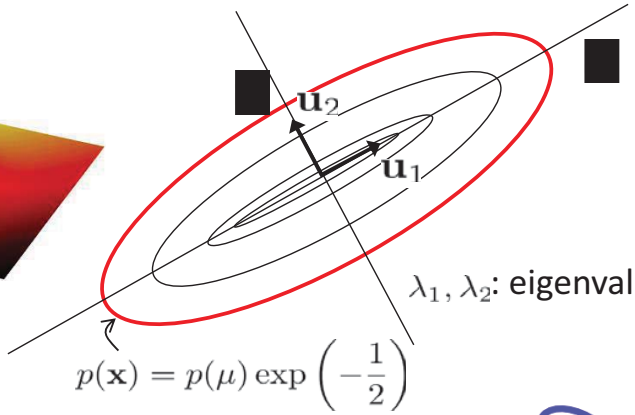
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

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

Principal axes are the eigenvector directions of Σ

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





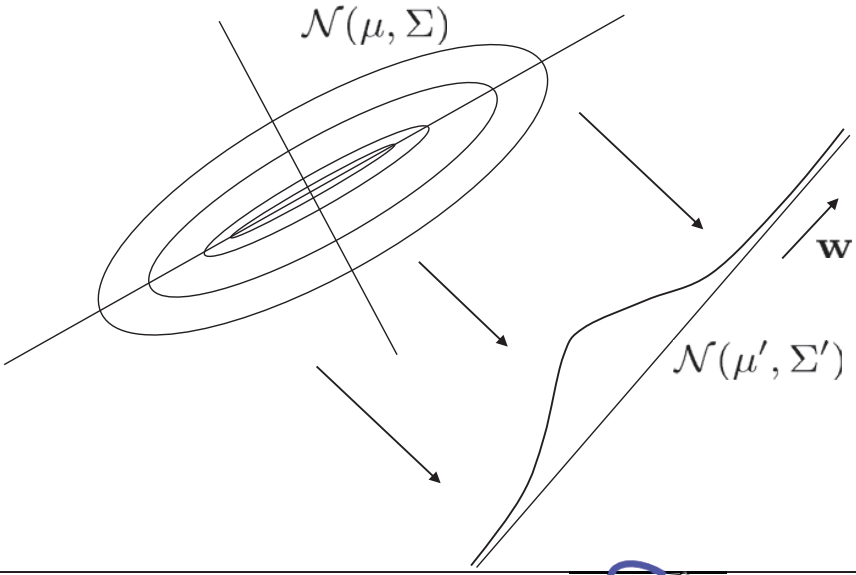
Gaussian Random Variable - Projection

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

Projection to any direction is Gaussian.

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

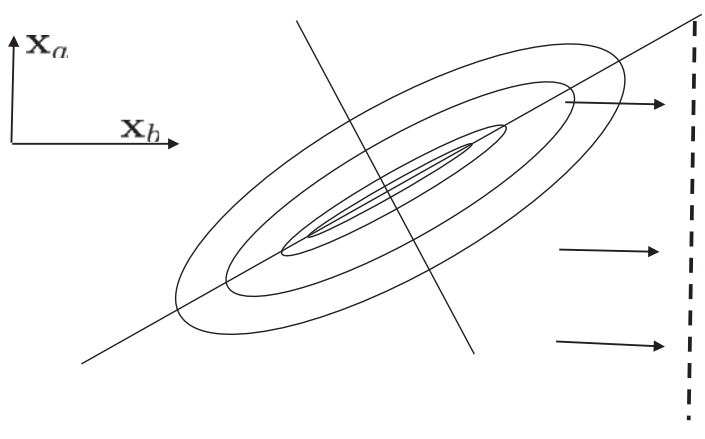
$$\Sigma' = \mathbf{w}^{\top}\Sigma\mathbf{w}$$



Gaussian Random Variable – Marginal

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

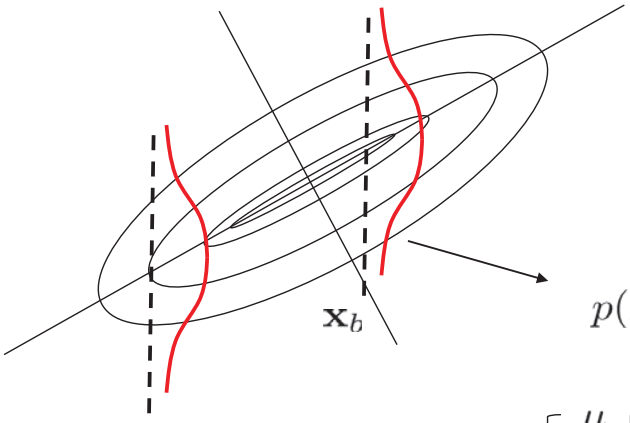
$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a \\ \mathbf{x}_b\end{pmatrix} \quad \begin{matrix}\mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b}\end{matrix} \qquad \mu = \begin{pmatrix}\mu_a \\ \mu_b\end{pmatrix} \quad \Sigma = \begin{pmatrix}\Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} &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



$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$
$$= \mathcal{N}(\mu_a, \Sigma_a)$$

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 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

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

where

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

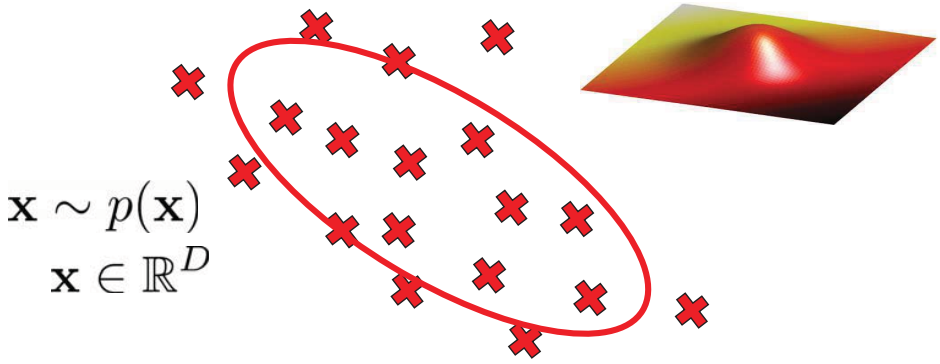
- For a new datum $\mathbf{x}$ with missing $\mathbf{x}_b$,

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

Parameter Estimation

Motivation – Parameter Estimation

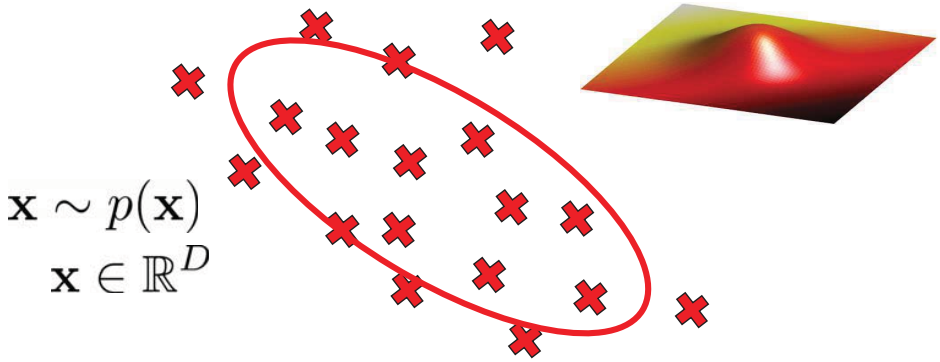
- Parameter estimation is an optimization problem



$\widehat{p}(\mathbf{x})$: 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

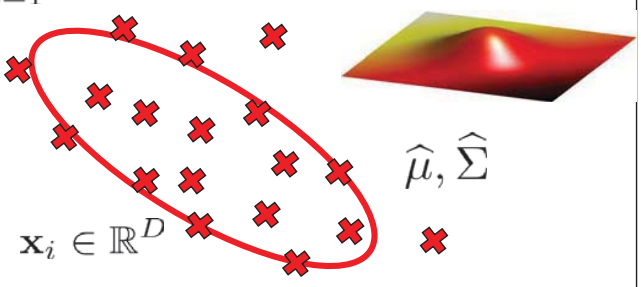
$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

- 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

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

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

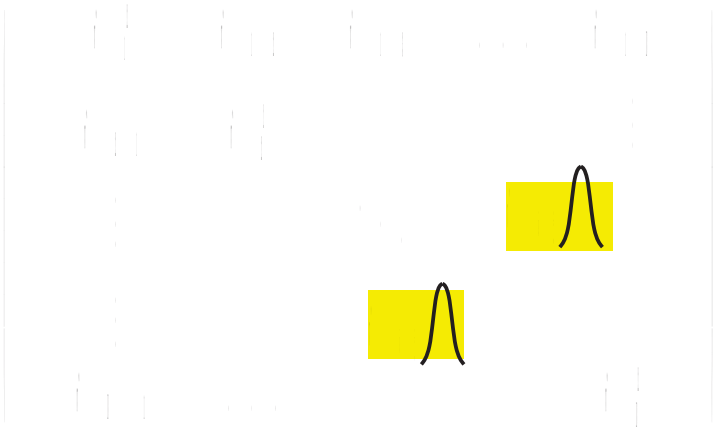
$$\frac{\partial \ln p(X|\mu,\Sigma)}{\partial \mu} = 0 \Rightarrow \widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

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

Covariance Estimation

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

- In high-dimensional space



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

Maximum A Posteriori (MAP) Estimation

- MAP estimation

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

- 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

$$\prod_{i=1}^{N} p(x_i|\mu)p(\mu) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right] \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

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

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

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

Maximum A Posteriori (MAP) Estimation for Gaussian

- Posterior density

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

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

- MAP of μ = Mean of μ = μ_n

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

MLE vs. MAP

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

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

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

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

MLE vs. MAP

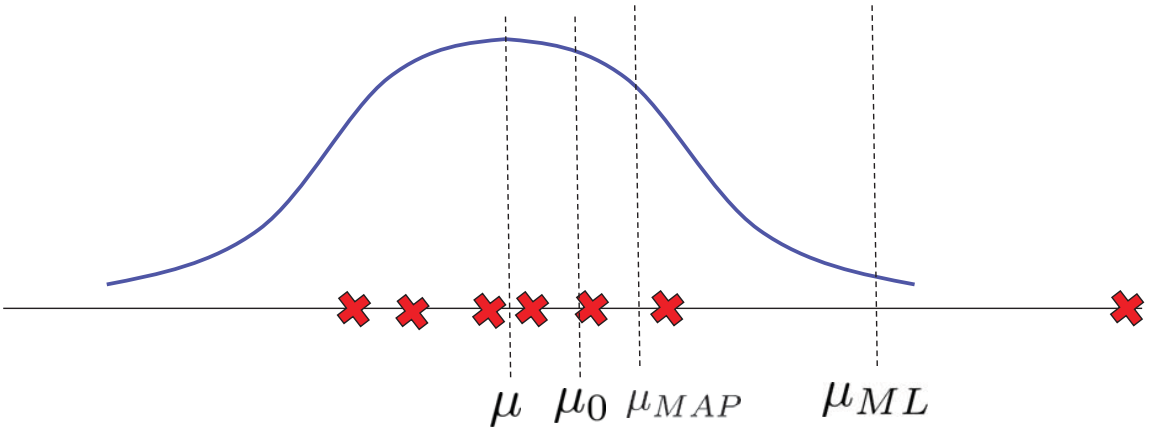
- For Gaussian
  - When we have a few outliers

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

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

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

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 the point estimate $\widehat{\mu}_{MAP}$ directly, but marginalize.

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

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

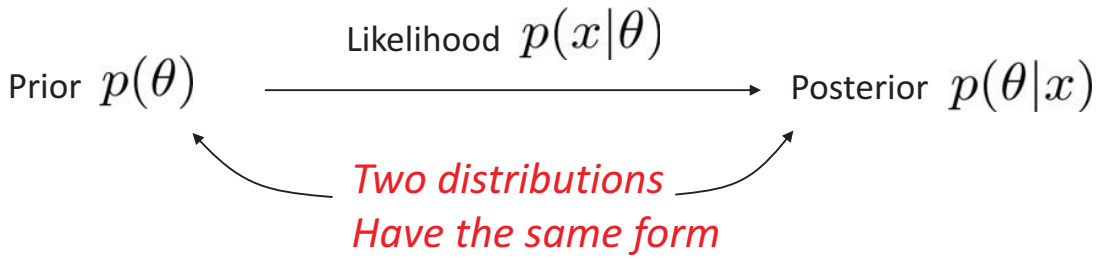
$$= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_n^2)}} \exp\left(-\frac{1}{2(\sigma^2+\sigma_n^2)}(x-\mu)^2\right)$$

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

Uncertainty of μ

Conjugate Priors

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



Conjugate Priors

Gaussian	Gaussian (μ)	Gaussian (μ)
Dirichlet	Multinomial	Dirichlet
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Gamma $(\lambda = \frac{1}{\sigma^2})$	Gaussian (σ^2)	Gamma $(\lambda = \frac{1}{\sigma^2})$

Kullback-Leibler Divergence

$$KL(p_e||p_\theta) = -\int p_e \log \frac{p_\theta}{p_e} d\mathbf{x}$$

p_e: Empirical density function

p_θ: Model density function

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

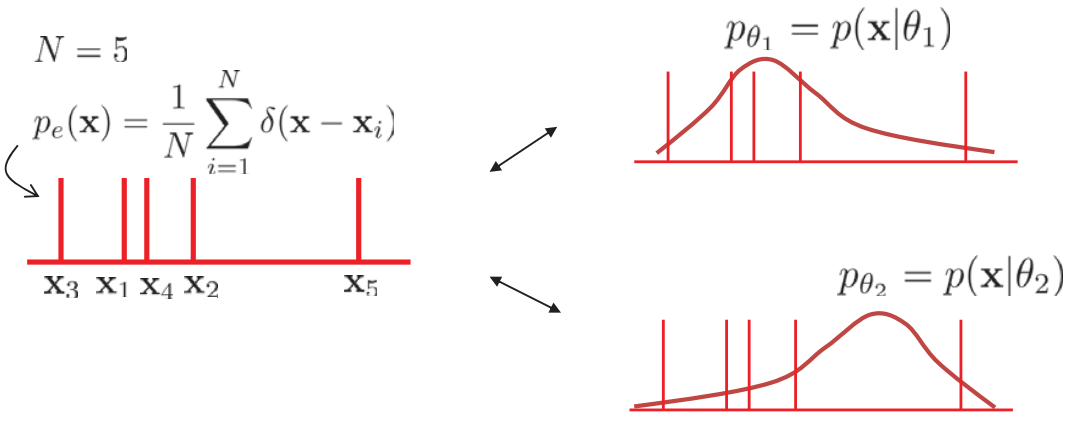
$$p_e = \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i)$$

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

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

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

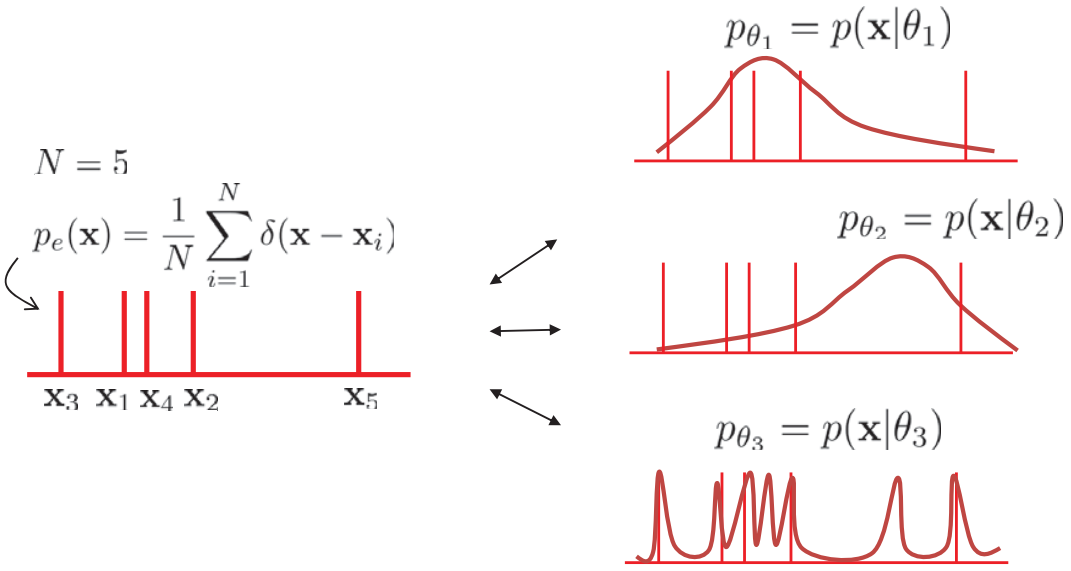
Kullback-Leibler Divergence



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