Introduction to Bayes Decision Theory

Heung-Il Suk

hisuk@korea.ac.kr

http://www.ku-milab.org

Department of Brain and Cognitive Engineering,
Korea University

June 19, 2018

Credit Scoring Problem as Classification

- Bank customer: {high-risk (1), low-risk (0)}

$$\mathbf{x} = \begin{bmatrix} x_1(\text{yearly income}) \\ x_2(\text{savings}) \end{bmatrix}$$

  - Credibility of a customer: Bernoulli random variable C conditioned on $[x_1, x_2]^\top$

$$\text{Decision} \begin{cases} C = 1 & \text{if } P(C=1|\mathbf{x}) > P(C=0|\mathbf{x}) \\ C = 0 & \text{otherwise} \end{cases}$$

  - Probability error

$$1 - \max\left[P(C=1|\mathbf{x}), P(C=0|\mathbf{x})\right]$$

Bayes' rule

$$
\begin{aligned}
P(C|\mathbf{x}) &= \frac{p(\mathbf{x}|C)P(C)}{p(\mathbf{x})} \\
&= \frac{p(\mathbf{x}|C)P(C)}{p(\mathbf{x}|C=1)P(C=1) + p(\mathbf{x}|C=0)P(C=0)}
\end{aligned}
$$

- $P(C)$: prior probability that $C = 1$ (regardless of $\mathbf{x}$)
- $p(\mathbf{x}|C)$: (class) likelihood, conditional probability that an event belonging to C has the associated observation $\mathbf{x}$
  - $p(\mathbf{x}|C=1)$: probability that a high-risk customer has $\mathbf{x}$
  - What the data tells us regarding the class
- $p(\mathbf{x})$: evidence, marginal probability that an observation $\mathbf{x}$ is seen, regardless of C

Bayes' rule (cont.)

$$
\begin{aligned}
P(C|\mathbf{x}) &= \frac{p(\mathbf{x}|C)P(C)}{p(\mathbf{x})} \\
&= \frac{p(\mathbf{x}|C)P(C)}{p(\mathbf{x}|C=1)P(C=1) + p(\mathbf{x}|C=0)P(C=0)}
\end{aligned}
$$

- $P(C|\mathbf{x})$: posterior probability

$$
\text{Decision rule: } \begin{cases} C = 1 & \text{if } P(C=1|\mathbf{x}) > P(C=0|\mathbf{x}) \\ C = 0 & \text{otherwise} \end{cases}
$$

K mutually exclusive and exhaustive classes

$$P(C_k) \geq 0 \text{ and } \sum_k P(C_k) = 1$$

$$P(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)P(C_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_k)P(C_k)}{\sum_{C_i} p(\mathbf{x}|C_i)P(C_i)}$$

Baye's classifier: for minimum error

$$\hat{k} = \operatorname*{argmax}_k P(C_k|\mathbf{x}) \qquad k \in \{1, \ldots, K\}$$

Loss and Risk

- In some cases, decisions are not equally good or costly
  - Loss for a high-risk applicant erroneously accepted vs. potential gain from a low-risk application erroneously rejected
  - Medical diagnosis
  - Earthquake prediction

Define loss λ_{ik} of action α_i when the real class is k

- Expected risk

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x})$$

  - Decision with a minimum risk

$$\hat{k} = \underset{k}{\text{argmin}}\, R(\alpha_k|\mathbf{x})$$

- e.g., 0/1 loss

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & i \neq j \end{cases}$$

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x}) = \sum_{k \neq i} P(C_k|\mathbf{x}) = 1 - P(C_i|\mathbf{x})$$

Machine Intelligence Lab

Reject: additional action α_{K+1}

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 0 < \lambda < 1 & \text{if } i = K+1 \\ 1 & \text{otherwise} \end{cases}$$

$$\begin{aligned} R(\alpha_{K+1}|\mathbf{x}) &= \sum_{k=1}^{K} \lambda P(C_k|\mathbf{x}) \\ &= ? \end{aligned}$$

$$R(\alpha_i|\mathbf{x}) = \sum_{k \neq i} P(C_k|\mathbf{x}) = ?$$

Machine Intelligence Lab

Optimal decision rule:

$$\hat{k} = \underset{k}{\operatorname{argmin}}\, R(\alpha_k|\mathbf{x}) \qquad k \in \{\underbrace{1, \ldots, K}_{\text{classes}}, \underbrace{K+1}_{\text{reject}}\}$$

$$\begin{cases} \text{Choose } C_k & \text{if} \quad P(C_k|\mathbf{x}) > P(C_i|\mathbf{x}) \quad (\forall k \neq i) \ \& \ P(C_k|\mathbf{x}) > 1 - \lambda \\ \text{Reject} & \text{otherwise} \end{cases}$$

- What if $\begin{cases} \lambda = 0 \\ \lambda \geq 1 \end{cases}$?

Interim Summary

- Classification: implementing a set of discriminant functions $g_k(\mathbf{x})$

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, g_k(\mathbf{x}) \qquad (k = 1, \ldots, K)$$

$$g_k(\mathbf{x}) \triangleq \begin{cases} -R(\alpha_k|\mathbf{x}) & \text{(Bayes' classifier)} \\ P(C_k|\mathbf{x}) & \text{(0/1 loss function)} \\ p(\mathbf{x}|C_k)P(C_k) & \text{(ignoring evidence term)} \end{cases}$$

Overview

$$P(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \times P(C)}{p(\mathbf{x})}$$

1. **Density estimation**: estimate the parameters of the distribution from the given samples
2. Plug in the estimates to the assumed model
3. Use the estimated distribution to make a decision (in combination with Bayes decision theory)

Density Estimation

To model $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations

- Fundamentally ill-posed: because there are infinitely many probability distributions that could have given rise to the observed finite data set

Parametric

- governed by a small number of adaptive parameters (*e.g.*, mean & variance in Gaussian)
- a procedure for determining values for the parameters
- Frequentist vs. Bayesian

Non-parametric

- the distribution form depends on the size of dataset
- parameters: control the model complexity, rather than the form
- histograms, nearest-neighbors, **kernels**

Frequentist

- Probability is frequency of a random, repeatable event
- Frequency of a tossed coin coming up heads is $1/2$

Bayesian

- Probability is a quantification of uncertainty
- Examples of uncertain events as probabilities
  - Whether the moon was once in its own orbit around the sun
  - Whether the Arctic ice cap will have disappeared by the end of the century

Machine Intelligence Lab

Gaussian Distribution

Univariate Gaussian Distribution

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

$\mu \in \mathbb{R}$: mean $\quad\quad \sigma^2 \in \mathbb{R}$: variance

Multivariate Gaussian Distribution

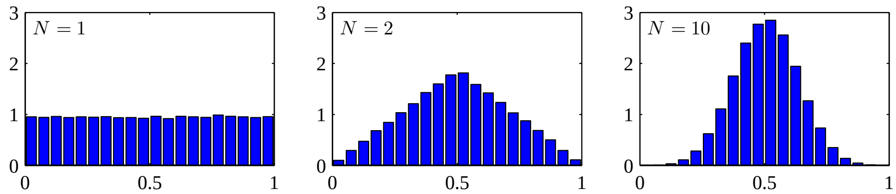
$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\Sigma\right) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$\boldsymbol{\mu} \in \mathbb{R}^D$: mean $\quad\quad \Sigma \in \mathbb{R}^{D\times D}$: covariance

Machine Intelligence Lab

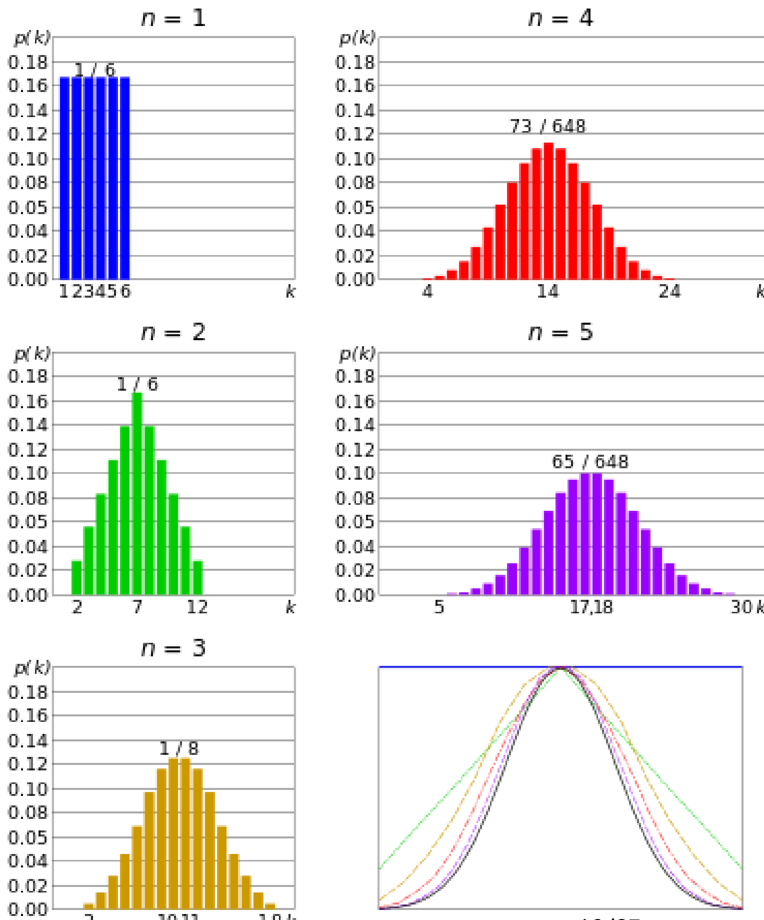
Why Gaussian Distribution in Various Contexts

- Central limit theorem: The sum of a set of random variables, which is itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.
  - In practice, the convergence to a Gaussian as N increases can be very rapid.

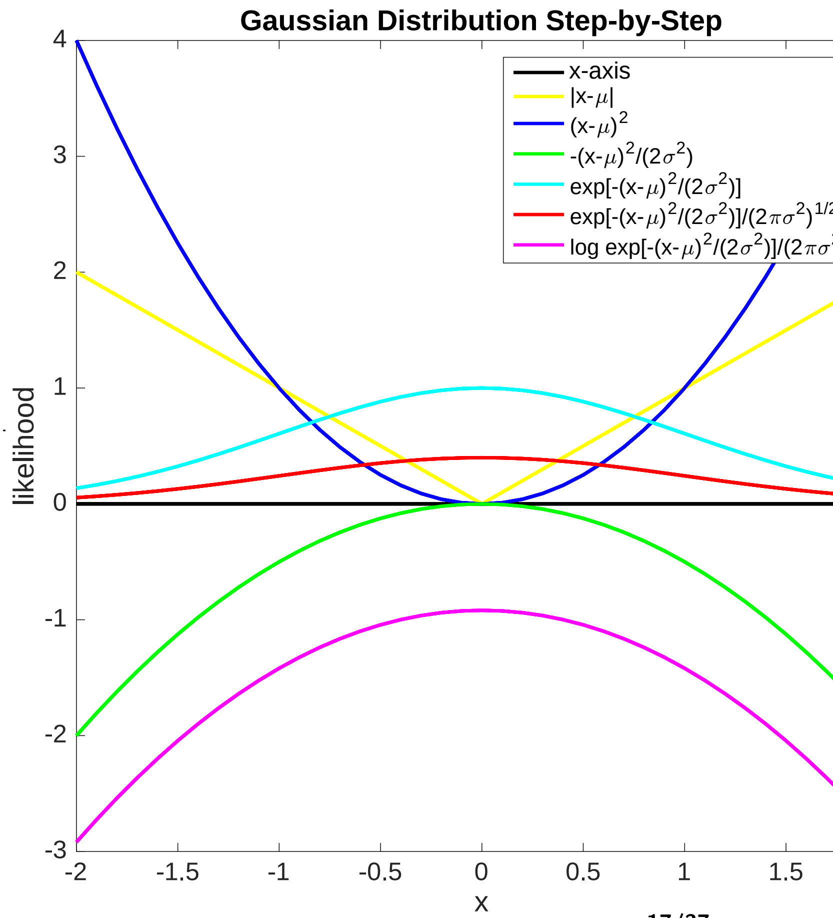


- Comparison of probability density functions, $p(k)$ for the sum of n fair 6-sided dice to show their convergence to a normal distribution with increasing n, in accordance to the central limit theorem.
- In the bottom-right graph, smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

[from Wikipedia]



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



Geometry of Multivariate Gaussian

Multivariate Gaussian Distribution

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

$\boldsymbol{\mu} \in \mathbb{R}^D$: mean $\qquad \Sigma \in \mathbb{R}^{D\times D}$: covariance

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}): \text{ Mahalanobis distance}$$

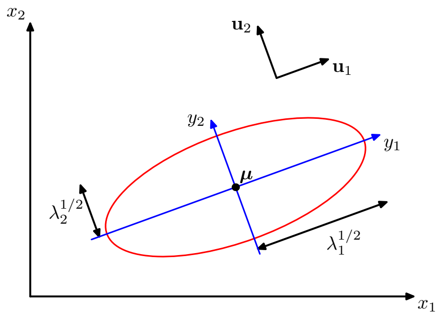
(Euclidean distance when $\Sigma = \mathbf{I}$)

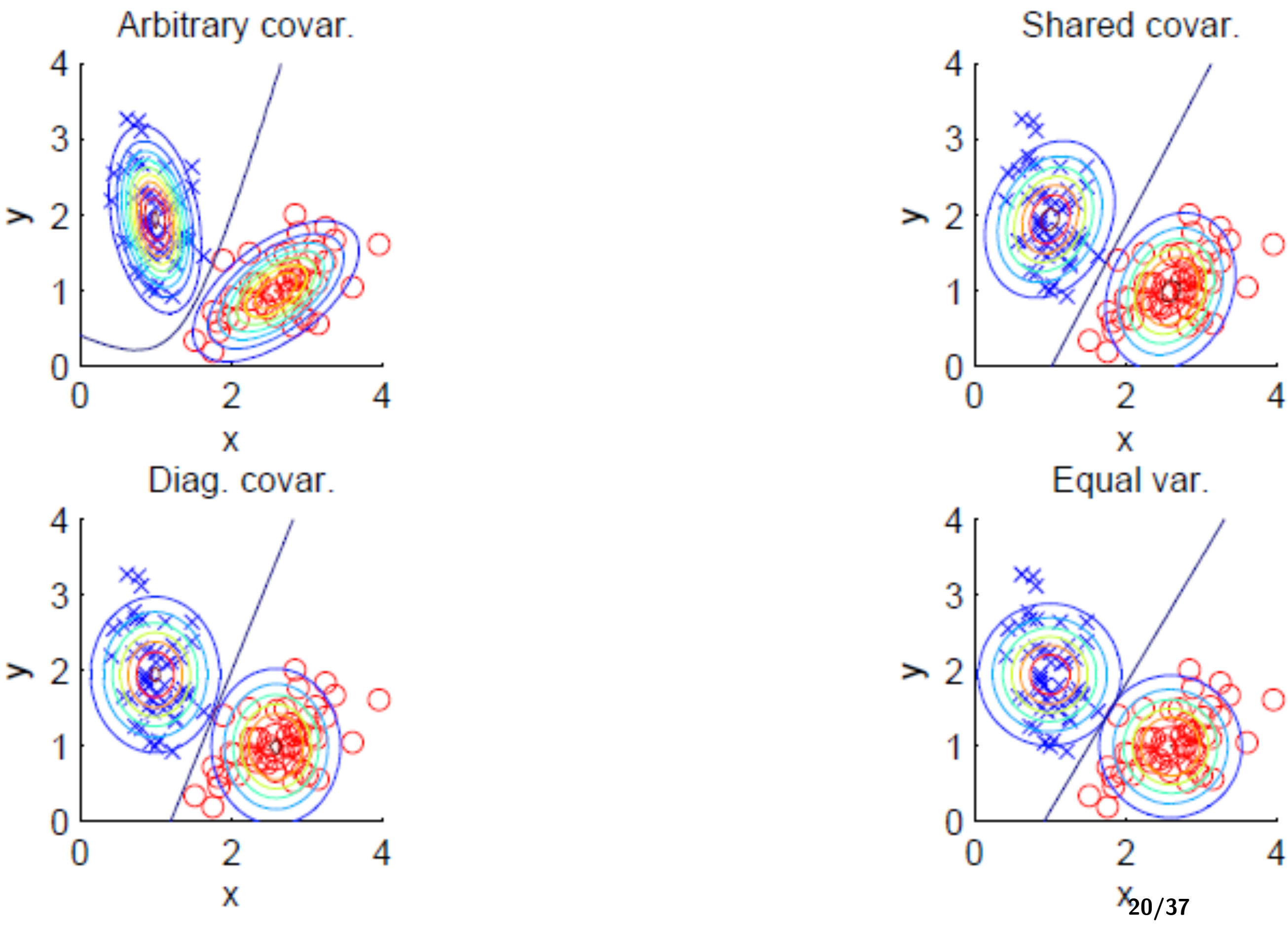
$$
\begin{aligned}
\Sigma \mathbf{u}_i &= \lambda_i \mathbf{u}_i (\text{eigendecomposition}) \\
\mathbf{u}_i^\top \mathbf{u}_j &= \mathbf{I}_{ij} (\mathbf{I}: \text{ identity matrix}) \\
\Sigma &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \\
\Sigma^{-1} &= \sum_{i=1}^{D} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top
\end{aligned}
$$

$$
\begin{aligned}
\Delta^2 &= (\mathbf{x}-\boldsymbol{\mu})^\top \left(\sum_{i=1}^{D} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top \right) (\mathbf{x}-\boldsymbol{\mu}) \\
&= \sum_i \lambda_i^{-1} (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{u}_i \mathbf{u}_i^\top (\mathbf{x}-\boldsymbol{\mu}) \\
&= \sum_i \lambda_i^{-1} \underbrace{\left\{ \mathbf{u}_i^\top (\mathbf{x}-\boldsymbol{\mu}) \right\}}_{\equiv d_i}^\top \left\{ \mathbf{u}_i^\top (\mathbf{x} -
\end{aligned}
$$

- Interpretation of $\{d_i\}$: a new coordinate system defined by the orthogonal vectors $\mathbf{u}_i$ that are shifted and rotated w.r.t. the original x_i coordinate

  $\mathbf{d} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) \quad \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_D]^\top$

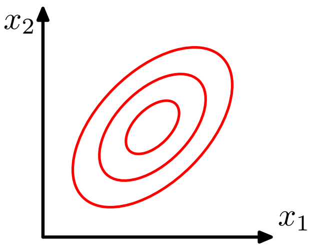




Limitations of Gaussian Distribution

- a large # of free parameters: $D\left(D+3\right)/2$
  - Too flexible in the sense of too many parameters
  - Restricting a covariance form: limits its ability to capture interesting correlations in the data

$$\begin{cases} \mathrm{diag}\left(\sigma_i^2\right): \text{ axis-aligned ellipsoid} \Rightarrow \#2D \\ \sigma^2\mathbf{I}: \text{ isotropic covariance} \Rightarrow \#\left(D+1\right) \end{cases}$$



(a) (b) (c)

Machine Intelligence Lab

Naïve Bayes

- Independent among variables, i.e., $\Sigma_{ij} = 0$ (for $i \neq j$)
- Mahalanobis distance $\rightarrow$ weighted $(1/\sigma_i)$ Euclidean distance

$$p(\mathbf{x}|C_i) = \prod_{d=1}^{D} p\left(x_d|C_i\right) = \frac{1}{(2\pi)^{D/2}\prod_{d=1}^{D}\sigma_d}\exp\left[-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_d-\mu_d}{\sigma_d}\right)^2\right]$$

- When $\sigma_d = \sigma$ (for $\forall d$) $\rightarrow$ Euclidean distance

Machine Intelligence Lab

Maximum Likelihood Estimation (MLE)

Given *i.i.d* samples $X = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ ($\mathbf{x}_n \sim p(\mathbf{x}|\theta)$),

- Find θ that makes sampling $\mathbf{x}_n$ from $p(\mathbf{x}|\theta)$ as likely as possible

$$l(\theta|X) \equiv p(X|\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta) \text{ (why?)}$$

Maximum Likelihood Estimation

$$L(\theta|X) \equiv \log l(\theta|X) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)$$

Parametric Classification

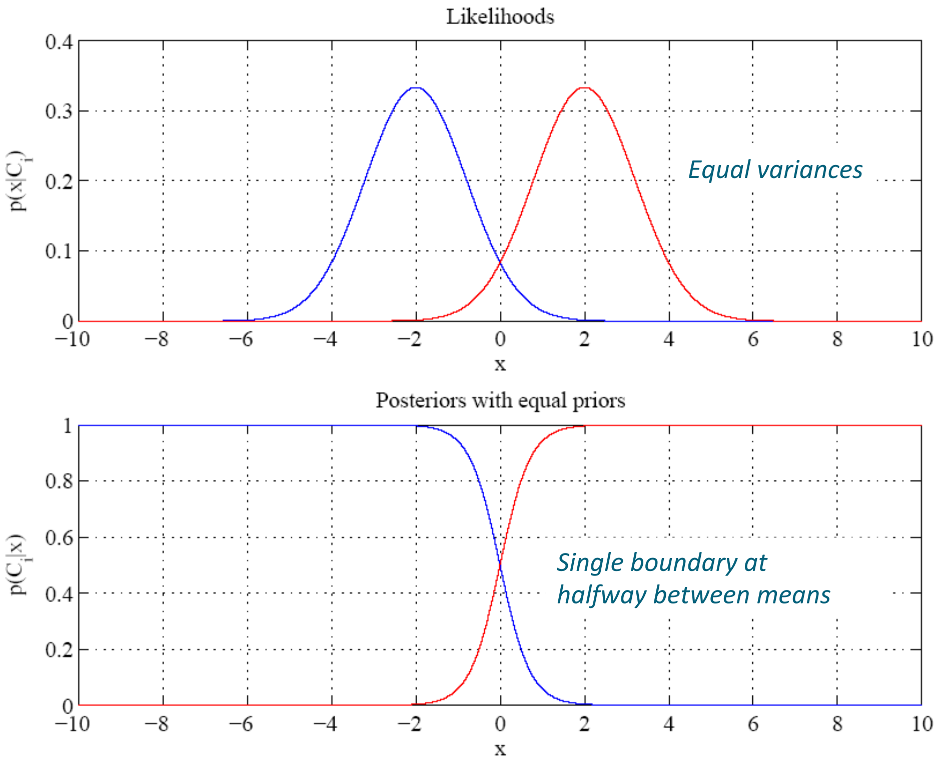
- Decision rule using a Bayes' rule: posterior prob. $P(C_i|\mathbf{x})$
- Discriminant function

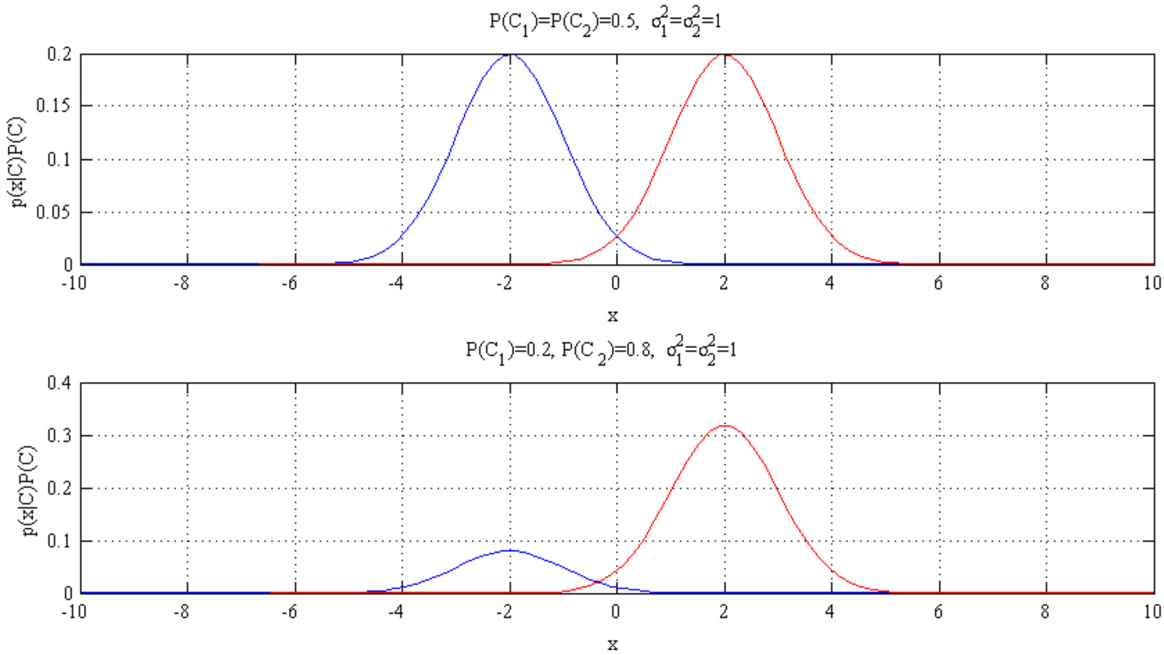
$$g_i(x) = p(\mathbf{x}|C_i)P(C_i)$$
$$g_i(x) = \log p(\mathbf{x}|C_i) + \log P(C_i)$$

  - Univariate Gaussian density

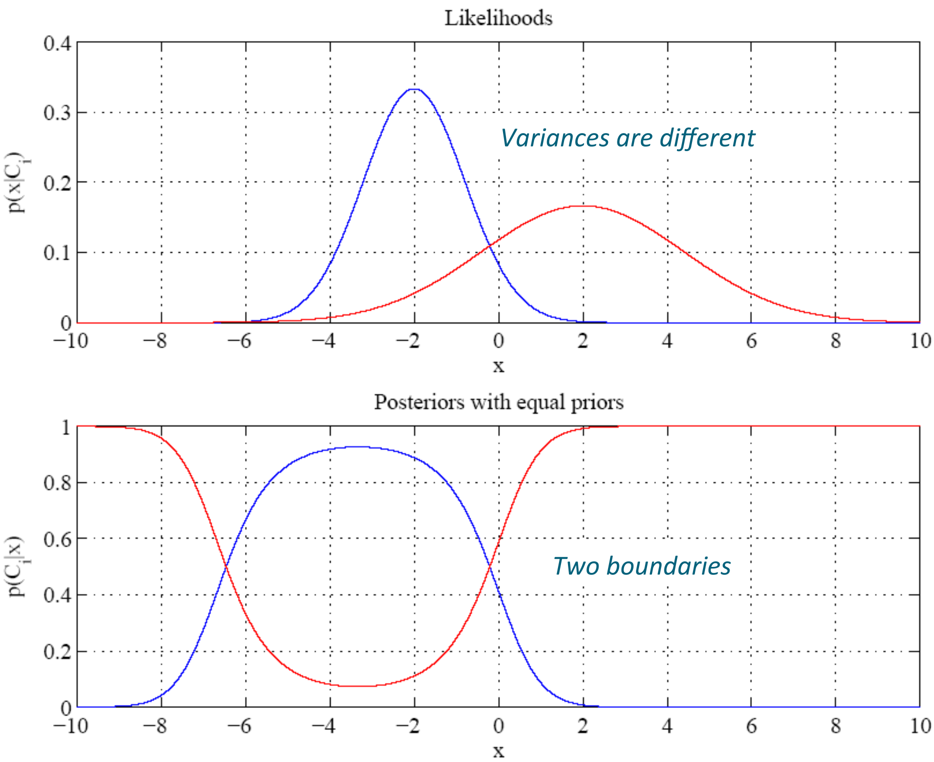
$$g_i(x) = -\log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

  - For equal prior and equal variance
    - $\hat{i} = \text{argmin}_i |x - m_i|$
    - Two-class boundary: $x = (m_1 + m_2)/2$





If priors are different, this has the effect of moving the threshold of decision
towards the mean of the less likely class.



Regression

$$y = f(\mathbf{x}) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Estimator $g(\mathbf{x}|\theta) \approx f(\mathbf{x})$

$$p(y|\mathbf{x}) \sim \mathcal{N}\left(g(\mathbf{x}|\theta), \sigma^2\right)$$

- Log likelihood

$$\begin{aligned} L(\theta|X) &= \log \prod_{n=1}^{N} p\left(\mathbf{x}_n, y_n\right) \\ &= \underbrace{\log \prod_{n=1}^{N} p\left(y_n|\mathbf{x}_n\right)}_{\text{Dependent on estimator}} + \underbrace{\log \prod_{n=1}^{N} p\left(\mathbf{x}_n\right)}_{\text{Independent of estimator}} \end{aligned}$$

$$
\begin{aligned}
L(\theta|X) &= \log \prod_{n=1}^{N} p\left(y_n|\mathbf{x}_n\right) \\
&= \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left\{y_n - g\left(\mathbf{x}_n|\theta\right)\right\}^2}{2\sigma^2}\right] \\
&= \log \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left\{y_n - g\left(\mathbf{x}_n|\theta\right)\right\}^2\right] \\
&= \underbrace{-N\log\left(\sqrt{2\pi}\sigma\right)}_{\text{Independent of parameter}} - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left\{y_n - g\left(\mathbf{x}_n|\theta\right)\right\}^2
\end{aligned}
$$

Machine Intelligence Lab

$$
\max L\left(\theta|X\right) = \min \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left\{y_n - g\left(\mathbf{x}_n|\theta\right)\right\}^2}_{E(\theta|X)}
$$

Least squares estimate

θ that minimizes $E\left(\theta|X\right)$

Machine Intelligence Lab

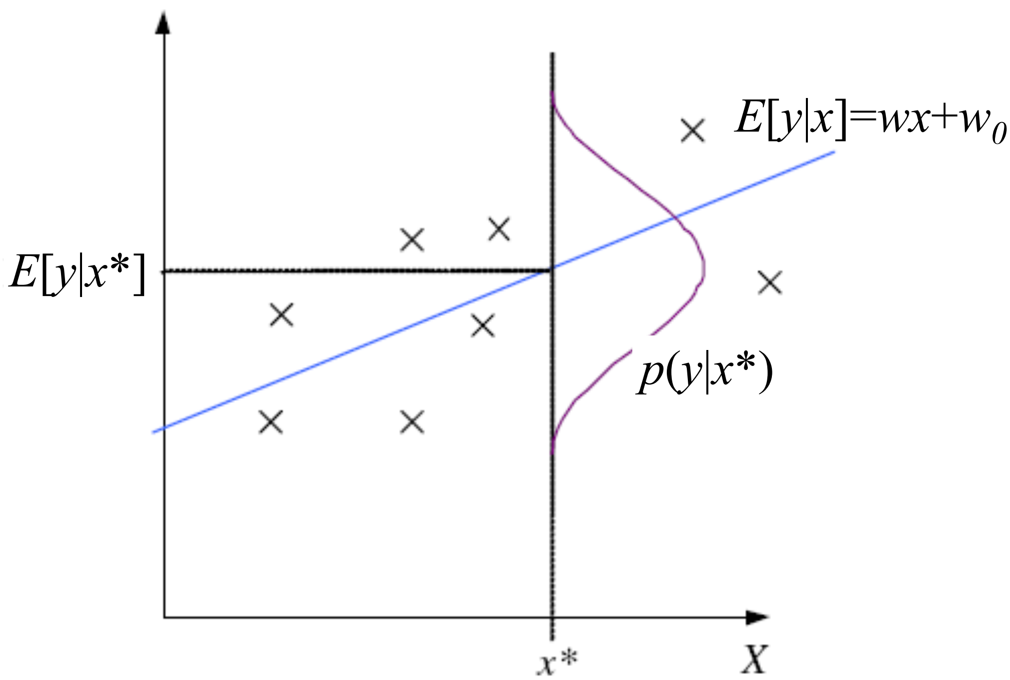
- Linear regression: $g\left(x_n|w_1, w_0\right) = w_1 x_n + w_0$

$$\frac{\partial E}{\partial w_0} \quad \Rightarrow \quad \sum_n y_n = N w_0 + w_1 \sum_n x_n$$

$$\frac{\partial E}{\partial w_1} \quad \Rightarrow \quad \sum_n y_n x_n = w_0 \sum_n x_n + w_1 \sum_n (x_n)^2$$

$$\underbrace{\begin{bmatrix} \sum_n y_n \\ \sum_n y_n x_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} N & \sum_n x_n \\ \sum_n x_n & \sum_n (x_n)^2 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\mathbf{w}}$$

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$$



MLE for Multivariate Gaussian

Given an i.i.d. dataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$,

- Log-likelihood function

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

  - depends on the dataset only through two quantities

$$\sum_{n=1}^{N}\mathbf{x}_n, \quad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top: \text{ `sufficient statistics'}$$

Machine Intelligence Lab

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

$$\frac{\partial}{\partial\Sigma}\ln p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = 0$$

$$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\top$$

Because of the joint maximization w.r.t. $\boldsymbol{\mu}$ and Σ, $\boldsymbol{\mu}_{\mathrm{ML}}$ is involved in Σ_{ML}.

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}; \qquad \mathbb{E}[\Sigma_{\mathrm{ML}}] = \frac{N-1}{N}\Sigma$$

Machine Intelligence Lab

- Discriminant function

$$g_i(\mathbf{x}) = \log p(\mathbf{x}|C_i) + \log P(C_i)$$

  - Multivariate Gaussian density: $p(\mathbf{x}|C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i)$$

    - $\mu_i \approx \mathbf{m}_i$ (sample mean)
    - $\Sigma_i \approx \mathbf{S}_i$ (sample covariance)
    - $\hat{P}(C_i) = \frac{\sum y_{n,i}}{N}$

Machine Intelligence Lab

Plugging estimates $\mathbf{m}_i$, $\mathbf{S}_i$, and $\hat{P}(C_i)$ into the discriminant function

$$\begin{aligned} g_i(x) &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^\top \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log\hat{P}(C_i) \\ &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}\left(\mathbf{x}^\top\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^\top\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^\top\mathbf{S}_i^{-1}\mathbf{m}_i\right) + \log\hat{P}(C_i) \\ &= \mathbf{x}^\top\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^\top\mathbf{x} + w_{i0} \quad (\textbf{quadratic discriminant}) \end{aligned}$$

$$\begin{aligned} \mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\ \mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\ w_{i0} &= -\frac{1}{2}\mathbf{m}_i^\top\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i) \end{aligned}$$

Machine Intelligence Lab

Further Issue with Gaussians

- Intrinsically unimodal (*i.e.*, having a single maximum)
  - Unable to provide a good approximation to multimodal distributions
  - Too limited in the range of distributions that it can adequately represent

- **Introducing latent variables**
  - Discrete latent variables: providing a rich family of multimodal distributions (*e.g.*, mixture of Gaussians)
  - Continuous latent variables: leading to models in which the number of free parameters can be controlled independently of the dimensionality D of the data space while still allowing the model to capture the dominant correlations in the data set
  - Discrete & continuous combined: extended to derive a very rich set of hierarchical models that can be adapted to a broad range of practical applications (*e.g.*, Gaussian version of Markov random field, linear dynamical system)

Thank you for your attention!!!

(Q & A)

hisuk (AT) korea.ac.kr

http://www.ku-milab.org