Lecture memo: Machine learning 1, Chapter 3

November 9, 2016

1 Introduction

For $y = \{w_1, w_2\}$, $\boldsymbol{x} \in \mathbb{R}^d$, Bayes decision theory

$$p(w_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|w_j)p(w_j)}{p(\boldsymbol{x})}$$

assumes complete knowledge of $p(w_j)$ and $p(\boldsymbol{x}|w_j)$. Typically, $p(w_j)$ and $p(\boldsymbol{x}|w_j)$ is unknown, and learned from samples (called training data).

In this chapter, you will learn

- How to learn $p(\boldsymbol{x}|w_j)$ from training data in supervised setting (estimating $p(\boldsymbol{x})$ is easy).
  - Maximum likelihood estimation
  - Maximum a posteriori estimation
  - Bayesian estimation
- Problem of dimenstionality
  - Computation
  - Overfitting

2 Assumptions

Labeled examples (supervised setting!) are given

$$\mathcal{D} = ((y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N))$$

$p(w_j|\mathcal{D}))$ can be estimated simply by counting the samples for specific data.

$$p(w_j|\mathcal{D}) \propto N_j = \sum_{n=1}^{N} \delta(y_i = w_j).$$

To estimate $p(\boldsymbol{x}|w_j)$, we separate the training data into c sets, according to the label.

$$\mathcal{D}_y = \{(y_i, \boldsymbol{x}_i); y_i = w_j).$$

If we focus on the data of a single class $w_j = 0$ or $w_j = 1$, the problem is density estimation:

$$\{\boldsymbol{x}_i; (y_i, \boldsymbol{x}_i) \in \mathcal{D}_y\} \sim p_j(\boldsymbol{x})$$

This is a very difficult problem. $p_j(\boldsymbol{x})$ is a function (infinite-dimensional vector!).

Typically, we assume a parametric form for $p_j(\boldsymbol{x})$, e.g., Gaussian

$$p(\boldsymbol{x}|\boldsymbol{\theta}_j) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}{2}\right)$$

Now, density estimation turned to parameter estimation$\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, of which the degree of freedom is finite $\boldsymbol{\mu}_j \in \mathbb{R}^M, \boldsymbol{\Sigma}_j \in \mathbb{S}^M_{++}$.

3 Maximum likelihood estimation

Omit the suffix j. Assume that data $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ are independently and identically *i.i.d.* distributed from $p(\boldsymbol{x}|\boldsymbol{\theta})$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

This is a density function of $\mathcal{D}$, but we see this as a function of $\boldsymbol{\theta}$ and call it likelihood.

Proposition 1 *(Likelihood principle) The model $p(\mathcal{D}|\boldsymbol{\theta})$ that best supports the observed data is more likely to be the true model than the ones that less supports.*

Maximum likelihood (ML) estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\mathcal{D}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(-\log p(\mathcal{D}|\boldsymbol{\theta}))$$

Negative log likelihood function

$$l(\boldsymbol{\theta}) = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

We solve the following problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}} l(\theta) \\ &\text{subject to} \qquad \boldsymbol{\theta} \in \Theta \end{aligned}$$

Θ is the domain of the parameter space.

If $l(\theta)$ is quasi-convex (or uni-modal) and differentiable, and no solution of the boundary of Θ, the stationary condition

$$\boldsymbol{\nabla} l(\theta) = \mathbf{0}$$

gives the global minimum. (ML estimator)

3.1 Gaussian case: $\boldsymbol{\mu}$ is unknown and $\boldsymbol{\Sigma}$ is known

Guassian

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

We solve the following problem

$$\begin{aligned}
&\text{Given } \boldsymbol{\Sigma} \in \mathbb{S}^d_{++} \\
&\min_{\boldsymbol{\mu}} l(\boldsymbol{\mu}) \\
&\text{subject to } \boldsymbol{\mu} \in \mathbb{R}^d
\end{aligned}$$

where

$$\begin{aligned}
l(\boldsymbol{\mu}) &= -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}) \\
&= \sum_{n=1}^{N} \frac{(\boldsymbol{\mu}-\boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{x}_n)}{2} + \text{const.}
\end{aligned}$$

This is a convex quadratic function! For

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n,$$

$$\begin{aligned}
l(\boldsymbol{\mu}) &= \sum_{n=1}^{N} \frac{(\boldsymbol{\mu}-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{x}_n)}{2} + \text{const.} \\
&= \sum_{n=1}^{N} \frac{(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}}) + 2(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{x}_n) + (\overline{\boldsymbol{x}}-\boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{x}_n)}{2} + \text{const.} \\
&= \frac{N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}})}{2} + \text{const.}
\end{aligned}$$

So,

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}$$

The ML estimator is the sample mean. The stationary condition also gives this solution

$$\mathbf{0} = \boldsymbol{\nabla} l(\theta) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{x}_n) = N\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \overline{\boldsymbol{x}})$$

3.2 Gaussian case: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown

$$\begin{aligned} &\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\text{subject to } \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{S}^d_{++} \end{aligned}$$

where

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}) \\ &= \frac{N}{2}\log|\boldsymbol{\Sigma}| + \sum_{n=1}^{N} \frac{(\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{x}_n)}{2} + \text{const.} \end{aligned}$$

3.2.1 1-dimensional case

$$l(\mu, \sigma^2) = \frac{N}{2}\log\sigma^2 + \sum_{n=1}^{N} \frac{(\mu - x_n)^2}{2\sigma^2} + \text{const.}$$

Stationary conditions

$$\begin{aligned} 0 &= \frac{\partial l}{\partial \mu} = \frac{(\mu - x_n)}{\sigma^2} \\ 0 &= \frac{\partial l}{\partial \sigma^2} = \frac{N}{2\sigma^2} - \sum_{n=1}^{N} \frac{(\mu - x_n)^2}{2\sigma^4} \end{aligned}$$

imply the ML estimator

$$\begin{aligned} \widehat{\mu} &= \overline{x} \\ \widehat{\sigma}^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{\mu})^2 \end{aligned}$$

3.2.2 General case

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{N}{2} \log |\boldsymbol{\Sigma}| + \sum_{n=1}^{N} \frac{(\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n)}{2} + \text{const.}$$

$$\begin{aligned}
\left(\frac{\partial l}{\partial \boldsymbol{\mu}}\right)_m &= \frac{\partial l}{\partial \mu_m} \\
\left(\frac{\partial l}{\partial \boldsymbol{\Sigma}}\right)_{l,m} &= \frac{\partial l}{\partial \Sigma_{l,m}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} (\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n) &= 2\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n), \\
\frac{\partial}{\partial \boldsymbol{\Sigma}} \log |\boldsymbol{\Sigma}| &= \boldsymbol{\Sigma}^{-1} \\
\frac{\partial}{\partial \boldsymbol{\Sigma}} \operatorname{tr}\left((\boldsymbol{\mu} - \boldsymbol{x}_n)(\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}\right) &= -\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n)(\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}
\end{aligned}$$

Using these,

$$\begin{aligned}
\boldsymbol{0} &= \frac{\partial l}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n), \\
\boldsymbol{0}_{d \times d} &= \frac{\partial l}{\partial \boldsymbol{\Sigma}} = \frac{N}{2} \boldsymbol{\Sigma}^{-1} - \sum_{n=1}^{N} \frac{\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}_n)(\boldsymbol{\mu} - \boldsymbol{x}_n)^\top \boldsymbol{\Sigma}^{-1}}{2}
\end{aligned}$$

Thus,

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \overline{\boldsymbol{x}}, \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})^\top
\end{aligned}$$

3.3 Bias

Definition 1 *An estimator $\boldsymbol{f}(\mathcal{D})$ is called an unbiased estimator for $\boldsymbol{\theta}$ if $\langle \boldsymbol{f}(\mathcal{D}) \rangle_{p(\mathcal{D}|\boldsymbol{\theta})} = \boldsymbol{\theta}$.*

$\widehat{\boldsymbol{\mu}}$ is an unbiased estimator, but $\widehat{\boldsymbol{\Sigma}}$ is not. An unbiased estimator is

$$\widehat{\sigma}^{2\ \mathrm{UB}} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \widehat{\mu})^2$$

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{UB}} = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \widehat{\boldsymbol{\mu}})^{\top}$$

Detailed will be in Chapter 9.

Definition 2 *An estimator $\boldsymbol{f}(\mathcal{D})$ is called an asymmetrically unbiased estimator for $\boldsymbol{\theta}$ if $\lim_{N\to\infty}\langle \boldsymbol{f}(\mathcal{D})\rangle_{p(\mathcal{D}|\boldsymbol{\theta})} = \boldsymbol{\theta}$.*

The ML estimators are asymmetrically unbiased.

4 Bayesian Estimation

To feedback the estimated result to the Bayes decision theory

$$p(w_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|w_j)p(w_j)}{p(\boldsymbol{x})}$$

we just use the estimated densities based on the observed data $\mathcal{D}$.

$$p(w_j|\boldsymbol{x},\mathcal{D}) = \frac{p(\boldsymbol{x}|w_j,\mathcal{D})p(w_j|\mathcal{D})}{p(\boldsymbol{x}|\mathcal{D})} = \frac{p(\boldsymbol{x}|w_j,\mathcal{D})p(w_j|\mathcal{D})}{\sum_{j=1}^{C} p(\boldsymbol{x}|w_j,\mathcal{D})p(w_j|\mathcal{D})}$$

In the case of ML estimation, we simply plug the ML estimator in

$$p(\boldsymbol{x}|w_j,\mathcal{D}_j) = p(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_j^{\mathrm{ML}})$$

A more principled way is to use the Bayes predictive distribution:

$$p(\boldsymbol{x}|w_j,\mathcal{D}_j) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}_j)d\boldsymbol{\theta}$$

$p(\boldsymbol{\theta}|\mathcal{D}_j)$ is a distribution of the parameter.

4.1 Bayesian Posterior and Predictive

4.1.1 1-D case (only σ^2 is known)

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

Prior distribution (our belief or prior guess before observation)

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

Bayes posterior is

$$\begin{aligned} p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu} \\ &\propto p(\mathcal{D}|\mu)p(\mu) \end{aligned}$$

$$p(\mu|\mathcal{D}) \propto \left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu - x_n)^2}{2\sigma^2}\right)\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Product of $N+1$ Gaussians! Gaussian is a member of exponential family, which is closed in multiplication (Any product of Gaussians is Gaussin).

Therefore the posterior is Gaussian!

$$\begin{aligned} p(\mu|\mathcal{D}) &\propto \left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu - x_n)^2}{2\sigma^2}\right)\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{\sigma^{-2}\sum_{n=1}^{N}(\mu - x_n)^2 + \sigma_0^{-2}(\mu-\mu_0)^2}{2}\right) \\ &\propto \exp\left(-\frac{N\sigma^{-2}\mu^2 - 2N\sigma^{-2}\overline{x}\mu + \sigma_0^{-2}\mu^2 - 2\sigma_0^{-2}\mu_0\mu}{2}\right) \\ &\propto \exp\left(-\frac{(N\sigma^{-2} + \sigma_0^{-2})\left(\mu - \frac{N\sigma^{-2}\overline{x} + \sigma_0^{-2}\mu_0}{N\sigma^{-2}+\sigma_0^{-2}}\right)^2}{2}\right) \end{aligned}$$

Therefore,

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

where

$$\begin{aligned} \widehat{\mu} &= \frac{N\sigma^{-2}\overline{x} + \sigma_0^{-2}\mu_0}{N\sigma^{-2} + \sigma_0^{-2}} \\ \widehat{\sigma}^2 &= \frac{1}{N\sigma^{-2} + \sigma_0^{-2}} \end{aligned}$$

Definition 3 *(conjugate prior) A prior $p(\boldsymbol{\theta})$ is said to be conjugate for a likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ if the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is in the same function class.*

With a conjugate prior, the posterior is called reproducing density. Conjugate prior with the function class arbitrary or one, of which expectation is hard to compute is useless!

$\widehat{\mu}$ is best guess after observation, and $\widehat{\sigma}^2$ indicates uncertainty. When $N \to \infty$, $\widehat{\mu} \to \overline{x}$, and $\widehat{\sigma}^2 \approx \frac{\sigma^2}{N}$. Prior affects the posterior only when we don't have enough observation.

The predictive distribution is given by

$$\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})d\mu \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} \exp\left(-\frac{(\mu-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right) d\mu \\
&\propto \int \exp\left(-\frac{\sigma^{-2}(x-\mu)^2 + \widehat{\sigma}^{-2}(\mu-\widehat{\mu})^2}{2}\right) d\mu \\
&= \int \exp\left(-\frac{\sigma^{-2}(x-\widehat{\mu}+\widehat{\mu}-\mu)^2 + \widehat{\sigma}^{-2}(\mu-\widehat{\mu})^2}{2}\right) d\mu \\
&= \exp\left(-\frac{\sigma^{-2}(x-\widehat{\mu})^2}{2}\right) \int \exp\left(-\frac{-2\sigma^{-2}(x-\widehat{\mu})(\mu-\widehat{\mu}) + \sigma^{-2}(\widehat{\mu}-\mu)^2 + \widehat{\sigma}^{-2}(\mu-\widehat{\mu})^2}{2}\right) d\mu \\
&= \exp\left(-\frac{\sigma^{-2}(x-\widehat{\mu})^2}{2}\right) \int \exp\left(-\frac{-2\sigma^{-2}(x-\widehat{\mu})\mu' + (\sigma^{-2}+\widehat{\sigma}^{-2})\mu'^2}{2}\right) d\mu' \\
&= \exp\left(-\frac{\sigma^{-2}(x-\widehat{\mu})^2 - \frac{\sigma^{-4}(x-\widehat{\mu})^2}{\sigma^{-2}+\widehat{\sigma}^{-2}}}{2}\right) \int \exp\left(-\frac{(\sigma^{-2}+\widehat{\sigma}^{-2})\left(\mu' - \frac{\sigma^{-2}(x-\widehat{\mu})}{\sigma^{-2}+\widehat{\sigma}^{-2}}\right)^2}{2}\right) d\mu'
\end{aligned}$$

where $\mu' = \mu - \widehat{\mu}$.

Thus,

$$\begin{aligned}
p(x|\mathcal{D}) &\propto \exp\left(-\frac{\sigma^{-4} + \sigma^{-2}\widehat{\sigma}^{-2} - \sigma^{-4}}{2(\sigma^{-2}+\widehat{\sigma}^{-2})}(x-\widehat{\mu})^2\right) \sqrt{2\pi(\sigma^{-2}+\widehat{\sigma}^{-2})^{-1}} \\
&\propto \exp\left(-\frac{\sigma^{-2}\widehat{\sigma}^{-2}}{2(\sigma^{-2}+\widehat{\sigma}^{-2})}(x-\widehat{\mu})^2\right)
\end{aligned}$$

and therefore

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

Note that

$$\lim_{N\to\infty} p(x|\mathcal{D}) = p(x|\widehat{\mu}^{\mathrm{ML}})$$

The bigger difference appears in more complex models, or hierarchical models. (when σ_0 is to be estimated.)

4.1.2 Multi-dimensional case ($\boldsymbol{\Sigma}$ is known)

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|_0^{1/2}} \exp\left(-\frac{(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right)$$

The posterior is

$$\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu})}{p(\mathcal{D})} \\
&\propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \\
&= \left(\prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\mu})\right) p(\boldsymbol{\mu}) \\
&= \left(\prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}_n-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})}{2}\right)\right) \\
&\qquad\qquad \cdot \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|_0^{1/2}} \exp\left(-\frac{(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right) \\
&\propto \left(\exp\left(-\frac{\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right)\right) \\
&\propto \left(\exp\left(-\frac{\sum_{n=1}^{N}(\boldsymbol{x}_n-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right)\right) \\
&\propto \left(\exp\left(-\frac{N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right)\right) \\
&\propto \left(\exp\left(-\frac{N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right)\right) \\
&\propto \left(\exp\left(-\frac{(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right)\right)
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= (N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}(N\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \\
\widehat{\boldsymbol{\Sigma}} &= (N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}
\end{aligned}$$

The predictive distribution is similarly computed as

$$\begin{aligned}
p(\boldsymbol{x}|\mathcal{D}) &= \int p(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\mathcal{D})d\boldsymbol{\mu} \\
&\propto \int \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right) d\boldsymbol{\mu} \\
&\propto \int \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}+\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}+\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right) d\boldsymbol{\mu}
\end{aligned}$$

$$
\begin{aligned}
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top}{2}\right)\\
&\qquad \int \exp\left(-\frac{-2(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})+(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})+(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right) d\boldsymbol{\mu}\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top}{2}\right)\\
&\qquad \int \exp\left(-\frac{-2(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})+(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top (\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right) d\boldsymbol{\mu}\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top}{2}\right)\\
&\qquad \int \exp\left(-\frac{-2(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'+\boldsymbol{\mu}'^\top (\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})\boldsymbol{\mu}'}{2}\right) d\boldsymbol{\mu}'\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top-(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)\\
&\int \exp\left(-\frac{(\boldsymbol{\mu}'-(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}))^\top(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{\mu}'-(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}))}{2}\right) d\boldsymbol{\mu}'
\end{aligned}
$$

where $\boldsymbol{\mu}' = \boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}$.

Thus,

$$
\begin{aligned}
p(\boldsymbol{x}|\mathcal{D}) &\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}-(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})-\boldsymbol{I})(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma}^{-1}+\widehat{\boldsymbol{\Sigma}}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)\\
&\propto \exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top (\boldsymbol{\Sigma}+\widehat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)
\end{aligned}
$$

and therefore

$$
p(\boldsymbol{x}|\mathcal{D}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}+\widehat{\boldsymbol{\Sigma}}|^{1/2}}\exp\left(-\frac{(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})^\top (\boldsymbol{\Sigma}+\widehat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}})}{2}\right)
$$

5 Recursive Bayesian approach

Bayes posterior

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}^N) &= \frac{p(\mathcal{D}^N|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}^N)} \\
&= \frac{p(\boldsymbol{\theta})\prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})\prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{p(\boldsymbol{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta})\prod_{n=1}^{N-1} p(\boldsymbol{x}_n|\boldsymbol{\theta})/p(\mathcal{D}^{N-1})}{\int p(\boldsymbol{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta})\prod_{n=1}^{N-1} p(\boldsymbol{x}_n|\boldsymbol{\theta})/p(\mathcal{D}^{N-1})d\boldsymbol{\theta}} \\
&= \frac{p(\boldsymbol{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{N-1})}{\int p(\boldsymbol{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{N-1})d\boldsymbol{\theta}}
\end{aligned}$$

Since $p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$, this gives a sequence $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}|\boldsymbol{x}_1)$, $p(\boldsymbol{\theta}|\{\boldsymbol{x}_1, \boldsymbol{x}_2\}), \ldots$. For simple distributions, we don't need to store all data $\mathcal{D}^N$ but only some statistics (mean and covariance in the case of Gaussian), which are called sufficient statistics.

6 ML vs. Bayes

- Computation (ML)
- Interpretability (ML)
- Accuracy (Bayes)

  More faithful. Predictive is not necessarily in the distribution class of the model especially when $p(\boldsymbol{\theta}|\mathcal{D})$ is asymmetric around $\widehat{\boldsymbol{\theta}}$. Also when we have hierarchical models (hyperparameter estimation).

7 Non-informative prior

Location parameter μ, scale parameter σ. Translation invariance $p(\mu) \propto 1$ which is improper. But approximated by Gaussian prior with large covariance.

Scale (unit like meters, fee, inches) invariance. $p(\log \sigma) \propto 1$. (scale appears as a shift.) Then,

$$p(\sigma)d\sigma = p(\log \sigma)d(\log \sigma) \propto d(\log \sigma) = \frac{1}{\sigma}d\sigma$$

Therefore, $p(\sigma) = \frac{1}{\sigma}$ which is also improper.

Uniform is parameter dependent. Uniform in σ is non-uniform in σ^2. Uniform in distribution based on the KL divergence is called Jafferys prior

$$p(\boldsymbol{\theta}) \propto \sqrt{|\boldsymbol{F}|}$$

where

$$F_{i,j} = \int \frac{\log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$$

7.1 Gibbs algorithm

Integration can be hard in general (depending on the distribution). Instead of using predictive

$$p(\boldsymbol{\theta}|\mathcal{D}) = \int p(x|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta},$$

we could draw a single sample from the posterior $\boldsymbol{\theta}' \sim p(\boldsymbol{\theta}|\mathcal{D})$ and use $p(x|\boldsymbol{\theta}')$. Under weak assumptions, the misclassification error is at most twice the expected error of Bayes optimal classifier. Note that this method is not what is called Gibbs sampling!

8 Problem of Dimensionality

Assume that $p(\boldsymbol{x}|w_j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), p(w_1) = p(w_2)$. Bayes error rate is

$$p(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} \exp(-u^2/2) du,$$

where r is the Mahalanobis distance

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

If the features are independent $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$,

$$r^2 = \sum_{i=1}^{d} \frac{(\mu_{i,1} - \mu_{i,2})^2}{\sigma_i^2}$$

The more d, the more r and smaller $p(e)$.

If you aren't satisfied with your classifier, you can add new features and increase d to make the classifier more discriminative.

This story only applies when we have accurate predictive $p(\boldsymbol{x}|w_j)$. If you estimate it, there is a trade-off. Also, there could be model error. $p(\boldsymbol{x}|w_j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ can be approximately correct, but rigorously wrong.

8.1 Computational complexity

8.1.1 Order of a function

$f(n) = O(h(n))$ is of the order of $O(h(n))$ (big oh of $f(n)$) if there exist constants c and n_0 such that $|f(n)| \leq c|h(n)|$ for all $n > n_0$. Namely, for a sufficiently large n, an upper bound on the function grows no worse than $h(n)$. For example, if $f(n) = n^4 + 3n + 1$, we can say $f(n) = O(n^4)$. ($c = 5, n_0 = 1$ or $c = 1.5, n_0 = 2, \ldots$ satisfy the condition) $O(\cdot)$ is not unique since if $f(n) = O(n^4)$, it holds that $f(n) = O(n^5)$, $f(n) = O(n^8)$, $f(n) = O(n^5 \log n)$. This is just an upper bound.

A unique version is big theta. $f(n)$ is of the order of $\Theta(h(n))$ (big theta of $f(n)$) if there exist constants c_1, c_2 and n_0 such that $c_1 h(n) \le f(n) \le c_2 h(n)$ for all $n > n_0$.

When considering computational complexity, we count the number of basic mathematical operations such as additions, multiplications, and divisions or time and memory in computer.

Computational complexity of our ML classifier with d dimension, N training samples, and c categories. In the training phase, computations are

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &: O(Nd) \\ \widehat{\boldsymbol{\Sigma}} &: O(Nd^2) \\ \widehat{\boldsymbol{\Sigma}}^{-1} &: O(d^3) \\ |\widehat{\boldsymbol{\Sigma}}| &: O(d^2) \end{aligned}$$

Since normally $N > d$, the biggest cost is $O(cNd^2)$.

In the test phase, we have to compute $(\boldsymbol{x} - \widehat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}})$ which costs $O(d)$ for $\boldsymbol{x} - \widehat{\boldsymbol{\mu}}$ and $O(d^2)$ for $\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}})$. Thus, the cost is $O(cd^2)$.

Notes: no info. for constants, memory requirement, parallelizability can be important. Exponential cost $O(2^N)$, $O(2^d)$ is regarded as infeasible if the variable is not very small.

8.2 Overfitting

If the degree of freedom is large compared to N, the estimation is poor (show regression case).

In this case, we can

- Reconsider features
- Do dimensionality reduction
- Use a common $\boldsymbol{\Sigma}$ for all classes
- Use $\lambda \boldsymbol{I} + (1 - \lambda)\widehat{\boldsymbol{\Sigma}}$. (original regularization)
- Use diagonal $\boldsymbol{\Sigma}$