

Pattern Recognition IV

Machine Learning

Michal Haindl

Faculty of Information Technology, KTI
Czech Technical University in Prague
Institute of Information Theory and Automation
Czech Academy of Sciences Prague, Czechia



<http://ro.utia.cz/>

MI/MIE-ROZ 2023-2024



Outline

- 1 Machine Learning
 - Transfer Learning
 - Learning Fairness
- 2 Statistical Learning
 - Convergence Concepts
 - Estimator Properties
- 3 Supervised Learning
 - Moment Method
 - Minimum χ^2
 - Maximum Likelihood
 - Bayesian Estimation
- 4 Nonparametric Distributions
 - Histogram Approach
 - Parzen Estimation



Outline

- 5 Unsupervised Learning
- 6 EM Algorithm
- 7 Supervised Versus Unsupervised Learning
- 8 Training Set Selection



Machine Learning

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed. (A. Samuel 1959)
- Building a model from an example training set of input observations in order to make data-driven predictions or decisions.



Machine Learning

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed. (A. Samuel 1959)
- Building a **model from an example training set** of input observations in order to make data-driven predictions or decisions.



Machine Learning

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed. (A. Samuel 1959)
- Building a **model from an example training set** of input observations in order to make data-driven predictions or decisions.

mathematical optimization / data mining / statistical learning /
computational learning / transfer learning / multi-task learning /
artificial intelligence. . .



Machine Learning

deep learning deep NN learning

statistical learning random variable / constants / models

deterministic learning parameters in deterministic models

transfer learning learning on one problem and applying it to a different but related problem

multi-task learning multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks

federated learning bringing code to the sensitive data owners and only sharing intermediate model training updates among them



Machine Learning

current methods rely on

- ⊖ huge amounts of training data
- ⊖ data are subject to measurement errors
- ⊖ vast computational resources
- ⊖ time-consuming trial and error methodology
- ⊖ learned results are difficult to generalize to related problems
- ⊖ cannot be leveraged to learn more complex concept
- ⊖ merit function is often multimodal $\Rightarrow$ several minima



ML should:

- not be a black box
- computationally efficient
- reduce the human involvement (user-specific parameters, model structure, distance metric ...)
- transparent
- interpretable (based on a physical principle)
- understandable
- adaptive



Let X and Ω be the input and output spaces, respectively.

$T = \{(X_1, \omega_1), (X_2, \omega_2), \dots, (X_n, \omega_n)\}$ be the training data, $f \in \mathcal{H}$ represents the objective function with $\mathcal{H}$ being its hypothesis space, $l(., .)$ a specific loss function, then ML:

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l(f(X_i), \omega_i)$$



Maximum Likelihood:

$$\theta^* = \arg \min_{\theta} L(\theta | X_1, \dots, X_n) = \arg \min_{\theta} f_{\theta}(X_1, \dots, X_n)$$



Structural Risk Minimization:

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l(f(X_i), \omega_i) + \lambda R(f)$$



Machine Learning

Statistical ML:

$$P_{train}(X, \Omega) = P_{test}(X, \Omega)$$

in reality often ($\rightarrow$ transfer learning)

$$P_{train}(X, \Omega) \neq P_{test}(X, \Omega)$$



Parametric / Non-parametric Learning

Parametric (model-based) learning:

- 1 assumption about the functional form $f(X, \theta)$
- 2 θ parameters estimation from T
 - $\oplus$ simplifies estimation $f(X, \theta)$ to parameters θ
 - $\ominus$ if $f(\cdot)$ is very different from true $\tilde{f}(\cdot) \Rightarrow$ poor estimate



Parametric / Non-parametric Learning

Non-parametric learning:

- 1 no assumption about the functional form $f(X, \theta)$
- 2 $f(\cdot)$ estimation from T
 - $\oplus$ potential to accurately fit a wider range of possible $f(\cdot)$
 - $\ominus$ do not reduce $f(X, \theta)$ estimation to small number of parameters $\Rightarrow$ much larger $\text{card}\{T_{non-par}\} \gg \text{card}\{T_{par}\}$ is required



Statistical Learning

Frequentist (Classical) Approach

- Parameters are fixed but unknown **constants**.
- Probabilities are interpreted as long-run relative frequency.
- Statistical procedures are judged on the long run over an infinite number of hypothetical repetitions of the experiment.



Statistical Learning

Frequentist (Classical) Approach

- Parameters are fixed but unknown **constants**.
- Probabilities are interpreted as long-run relative frequency.
- Statistical procedures are judged on the long run over an infinite number of hypothetical repetitions of the experiment.

Bayesian Approach

- Parameters are **random variables**.
- Probability rules used for their inferences.
- Probability statements - degree of belief.
- Prior distribution & observed data $\rightarrow$ revised belief.



Statistical Learning

optimal Bayesian classifier design $P(\omega_i), p(X|\omega_i), K$

- 1 $P(\omega_i), p(X|\omega_i), K$ completely known
- 2 $P(\omega_i), T_i, K$ known, $p(X|\omega_i)$ known up to θ_i
- 3 $P(\omega_i), K, T_i^a, T_i^b$ known, $p(X|\omega_i)$ known up to θ_i
- 4 $P(\omega_i), T, K$ known, $p(X|\omega_i)$ known up to θ_i
- 5 $P(\omega_i), T$ known, $p(X|\omega_i)$ known up to θ_i
- 6 $P(\omega_i), p(X|\omega_i), K$ completely unknown



Statistical Learning

optimal Bayesian classifier design $P(\omega_i), p(X|\omega_i), K$

- 1 $P(\omega_i), p(X|\omega_i), K$ completely known **no learning**
- 2 $P(\omega_i), T_i, K$ known, $p(X|\omega_i)$ known up to θ_i **supervised**
- 3 $P(\omega_i), K, T_i^a, T_i^b$ known, $p(X|\omega_i)$ known up to θ_i **semi-sup.**
- 4 $P(\omega_i), T, K$ known, $p(X|\omega_i)$ known up to θ_i **unsupervised**
- 5 $P(\omega_i), T$ known, $p(X|\omega_i)$ known up to θ_i **unsupervised**
- 6 $P(\omega_i), p(X|\omega_i), K$ completely unknown **non-parametric unsupervised**



Identifiability

$\exists!$ solution if $p(X|\theta)$ is identifiable

pdf $p(X|\theta)$ is **identifiable** if $\theta \neq \theta'$ implies $\exists X : p(X|\theta) \neq p(X|\theta')$
 $p(X|\theta)$ **not identifiable** $\rightarrow$ unique θ cannot be recovered

- $\oplus$ most continuous pdf (Gaussian)
- $\ominus$ discrete pdf mixtures
- $\ominus$ too many components & few samples
- $\odot$ unsupervised learning is principally impossible for not identifiable pdf $\rightarrow$ **distinct classifications** for identical data
- $\odot$ no problem in modeling



Statistical Parameter Estimation

- histogram method
- moment method
- minimum χ^2
- maximum (pseudo-)likelihood
 - parameters unknown fixed quantities
 - maximize the probability of obtaining samples actually observed
- Bayesian estimation
 - parameters random variables
 - prior distribution
 - posterior density



Statistical Parameter Estimation

- histogram method
- moment method
- minimum χ^2
- maximum (pseudo-)likelihood
 - parameters unknown fixed quantities
 - maximize the probability of obtaining samples actually observed
- Bayesian estimation
 - parameters random variables
 - prior distribution
 - posterior density



Statistical Parameter Estimation

- histogram method
- moment method
- minimum χ^2
- maximum (pseudo-)likelihood
 - parameters unknown fixed quantities
 - maximize the probability of obtaining samples actually observed
- Bayesian estimation
 - parameters random variables
 - prior distribution
 - posterior density



Statistical Parameter Estimation

- histogram method
- moment method
- minimum χ^2
- maximum (pseudo-)likelihood
 - parameters unknown fixed quantities
 - maximize the probability of obtaining samples actually observed
- Bayesian estimation
 - parameters random variables
 - prior distribution
 - posterior density



Statistical Parameter Estimation

- histogram method
- moment method
- minimum χ^2
- maximum (pseudo-)likelihood
 - parameters unknown fixed quantities
 - maximize the probability of obtaining samples actually observed
- Bayesian estimation
 - parameters random variables
 - prior distribution
 - posterior density



Estimator Properties - biasness

frequentist criteria (in Bayesian statistics as a pre-posterior analysis)

- unbiased $\hat{\theta}$ if

$$E\{\hat{\theta}\} = \int \dots \int \hat{\theta} p(X_1, \dots, X_n | \theta) dX_1 \dots dX_n = \theta$$

bias function $b(\theta) = E\{\hat{\theta}\} - \theta$

- asymptotically unbiased $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} E\{\hat{\theta}\} = \theta$$

unbiased $\hat{\theta} \Rightarrow$ asymptotically unbiased
 asymptotically unbiased $\hat{\theta} \not\Rightarrow$ unbiased

- consistent $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$$

(convergence in probability)



Estimator Properties - biasness

- unbiased $\hat{\theta}$ if

$$E\{\hat{\theta}\} = \int \dots \int \hat{\theta} p(X_1, \dots, X_n | \theta) dX_1 \dots dX_n = \theta$$

bias function $b(\theta) = E\{\hat{\theta}\} - \theta$

- asymptotically unbiased $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} E\{\hat{\theta}\} = \theta$$

unbiased $\hat{\theta} \Rightarrow$ asymptotically unbiased
 asymptotically unbiased $\hat{\theta} \not\Rightarrow$ unbiased

- consistent $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$$

(convergence in probability)



Estimator Properties - biasness, consistency

- unbiased $\hat{\theta}$ if

$$E\{\hat{\theta}\} = \int \dots \int \hat{\theta} p(X_1, \dots, X_n | \theta) dX_1 \dots dX_n = \theta$$

bias function $b(\theta) = E\{\hat{\theta}\} - \theta$

- asymptotically unbiased $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} E\{\hat{\theta}\} = \theta$$

unbiased $\hat{\theta} \Rightarrow$ asymptotically unbiased
 asymptotically unbiased $\hat{\theta} \not\Rightarrow$ unbiased

- consistent $\hat{\theta}$ if

$$\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$$

(convergence in probability)



Estimator Properties - efficiency

- efficient $\hat{\theta}$ if

1 is unbiased

2 $E\{(\hat{\theta} - \theta)^2\} \geq \frac{1}{\sigma_n^2(\theta)}$ efficient if $E\{(\hat{\theta} - \theta)^2\} = \frac{1}{\sigma_n^2(\theta)}$

$$\begin{aligned}\sigma_n^2(\theta) &= E \left\{ \left(\frac{\partial}{\partial \theta} \log p(X_1, \dots, X_n | \theta) \right)^2 \right\} \text{ Fisher's information} \\ &= -E \left\{ \frac{\partial^2}{\partial \theta^2} \log p(X_1, \dots, X_n | \theta) \right\}\end{aligned}$$

(inv. of Cramer-Rao lower bound)



Estimator Properties - sufficiency

- sufficient $\hat{\theta}$ if for any other $\tilde{\theta}$

$$p(\tilde{\theta} | \hat{\theta}, \theta) = p(\tilde{\theta} | \hat{\theta})$$

i.e.

$$p(\theta | \hat{\theta}, \tilde{\theta}) = p(\theta | \hat{\theta})$$



Estimator Properties - sufficiency

- sufficient $\hat{\theta}$ if for any other $\tilde{\theta}$

$$p(\tilde{\theta} | \hat{\theta}, \theta) = p(\tilde{\theta} | \hat{\theta})$$

i.e.

$$p(\theta | \hat{\theta}, \tilde{\theta}) = p(\theta | \hat{\theta})$$

sufficient statistic s - a function of samples containing all the information necessary for θ estimation i.e. $p(\theta | s, T) = p(\theta | s)$



Estimator Properties - Bias x Precision



low bias, high precision



Estimator Properties - Bias x Precision



low bias, high precision



high bias, low variance



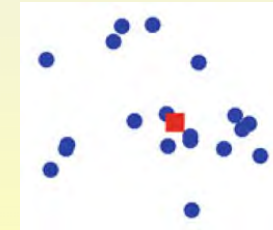
Estimator Properties - Bias x Precision



low bias, high precision



high bias, low variance



low bias, high variance



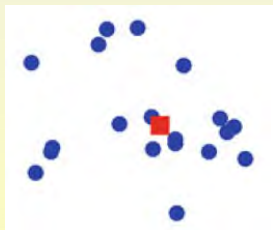
Estimator Properties - Bias x Precision



low bias, high precision



high bias, low variance



low bias, high variance



high bias and variance



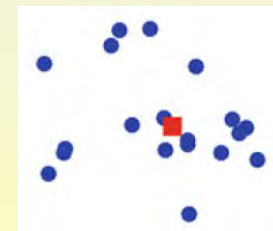
Estimator Properties - Bias x Precision



low bias, high precision



high bias, low variance



low bias, high variance



high bias and variance

high precision (low variance), low bias (high accuracy)



Supervised Learning - Moment Method

assumptions

- X_i conditional independent
- parametric form of $p(X|\omega_i)$ is known
- T_i gives no information about θ_j $j \neq i$, different classes are functionally independent $\rightarrow K$ problems

theoretical moments are compared with sampling moments

$\mu_i = E\{X_j^i\}$ theoretical moment

$M_i = \frac{1}{n} \sum_{j=1}^n X_j^i$ sampling moment

$\bar{M}_i = \frac{1}{n} \sum_{j=1}^n (X_j - M_1)^i$ central sampling moment

$$\mu_i = M_i \quad i = 1, 2, \dots$$



Moment Method 2

e.g. $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu_1 = M_1$$

$$\mu_2 = \bar{M}_2$$

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \bar{M}_2$$

$\hat{\sigma}^2$ is asymptotically unbiased, but biased



Moment Method 2

- ⊕ simple
- ⊕ consistent
- ⊖ biased $\Rightarrow$ not efficient
- ⊖ impossible for pdf with infinite moments (Cauchy)

Cauchy density

$$p(x; x_0, \gamma) = \frac{1}{\pi\gamma} \left[\frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \right]$$

x_0 location parameter, γ scale parameter

$$\mu_n = \int_{-\infty}^{\infty} \frac{\Gamma}{2\pi} \frac{x^n}{(x-m)^2 + (0.5\Gamma)^2} dx \quad \text{diverge for } n \geq 1$$

mean, variance - undefined

Minimum χ^2

assumptions

- $X \in$ continuous distribution
- observations X_i classified into K intervals with frequencies $n_1, \dots, n_K$
- $p_i(\theta)$ $i = 1, \dots, K$ their probabilities (function of θ)

n_i $i = 1, \dots, K$ observed frequencies

$E\{n_i\} = n p_i(\theta)$ $i = 1, \dots, K$ mean (hypothetical)

frequencies $n = \sum_i n_i$

- minimum χ^2

$$\min_{\theta} \chi^2 = \min_{\theta} \left\{ \sum_i \frac{[n_i - n p_i(\theta)]^2}{n p_i(\theta)} \right\}$$



Minimum χ^2 Methods

- minimum modified χ^2

$$\min_{\theta} \tilde{\chi}^2 = \min_{\theta} \left\{ \sum_i \frac{[n_i - n p_i(\theta)]^2}{n_i} \right\}$$

$$0 = \sum_i \frac{[n_i - n p_i(\theta)]}{n_i} \frac{\partial p_i(\theta)}{\partial \theta}$$

- minimum Kullback-Leibler separator

$$\min_{\theta} KL = \min_{\theta} \left\{ \sum_i p_i(\theta) \log \frac{p_i(\theta)}{\frac{n_i}{n}} \right\}$$

$$0 = \sum_i \left[\frac{n_i}{n} + \log \frac{n p_i(\theta)}{n_i} \right] \frac{\partial p_i(\theta)}{\partial \theta}$$

⊕ consistent

⊕ efficient



Maximum A Posteriori Estimation

$$\theta^{MAP} = \arg \max_{\theta} p(\theta | T_i)$$

Maximum likelihood:

$$\theta^{ML} = \arg \max_{\theta} p(T_i | \theta)$$

Maximum pseudo-likelihood:

$$\theta^{PML} = \arg \max_{\theta} \prod_{j=1}^{n_i} p(X_j | \theta)$$



Maximum Likelihood Estimation

assumptions

- parametric form of $p(X|\omega_i)$ is known
- X_i conditional independent (usually / pseudo-likelihood)
- T_i gives no information about θ_j $j \neq i$, different classes are functionally independent $\rightarrow K$ problems

likelihood of θ

$$p(T_i | \theta) = \prod_{j=1}^{n_i} p(X_j | \theta)$$

$$\hat{\theta} = \arg \max_{\theta} \{p(T_i | \theta)\}$$

log-likelihood

$$L(\theta) = \log p(T_i | \theta) = \sum_{j=1}^{n_i} \log p(X_j | \theta)$$

the maximum likelihood estimate $\nabla_{\theta} L = 0$



Maximum Likelihood Estimation

assumptions

- parametric form of $p(X|\omega_i)$ is known
- X_i conditional independent (usually / pseudo-likelihood)
- T_i gives no information about θ_j $j \neq i$, different classes are functionally independent $\rightarrow K$ problems

likelihood of θ

$$p(T_i | \theta) = \prod_{j=1}^{n_i} p(X_j | \theta)$$

$$\hat{\theta} = \arg \max_{\theta} \{p(T_i | \theta)\}$$

log-likelihood

$$L(\theta) = \log p(T_i | \theta) = \sum_{j=1}^{n_i} \log p(X_j | \theta)$$

the maximum likelihood estimate $\nabla_{\theta} L = 0$



Supervised Bayesian Estimation

assumptions

- 1 the form of cpdf $p(X|\theta)$ is known except the θ vector
- 2 prior pdf $p(\theta)$ is known
- 3 different classes are functionally independent
- 4 conditionally independent training samples $T_i = \{X_1, \dots, X_{n_i}\}$ are available $\downarrow$

$$p(X|T_i) = \int p(X|\theta)p(\theta|T_i)d\theta$$

$$p(\theta|T_i) = \frac{p(T_i|\theta)p(\theta)}{\int p(T_i|\theta)p(\theta)d\theta} \quad (1)$$



Bayesian Estimation 2

from assumption 4.

$$p(T_i|\theta) = \prod_{j=1}^{n_i} p(X_j|\theta)$$

$$p(\theta|T_i) \propto \prod_{j=1}^{n_i} p(X_j|\theta)p(\theta)$$

- Bayesian learning - recursive parameter estimation if $t > 1$

$$p({}^t T_i|\theta) = p({}^t X|\theta) p({}^{t-1} T_i|\theta)$$

$$p(\theta|{}^t T_i) = \frac{p({}^t X|\theta) p(\theta|{}^{t-1} T_i)}{\int p({}^t X|\theta) p(\theta|{}^{t-1} T_i)d\theta}$$

$$p(\theta|{}^0 T_i) = p(\theta)$$



Bayesian Estimation 2

from assumption 4.

$$p(T_i|\theta) = \prod_{j=1}^{n_i} p(X_j|\theta)$$

$$p(\theta|T_i) \propto \prod_{j=1}^{n_i} p(X_j|\theta)p(\theta)$$

- Bayesian learning - recursive parameter estimation if $t > 1$

$$p({}^t T_i|\theta) = p({}^t X|\theta) p({}^{t-1} T_i|\theta)$$

$$p(\theta|{}^t T_i) = \frac{p({}^t X|\theta) p(\theta|{}^{t-1} T_i)}{\int p({}^t X|\theta) p(\theta|{}^{t-1} T_i)d\theta}$$

$$p(\theta|{}^0 T_i) = p(\theta)$$



Bayesian Estimation 3

for usual pdf the sequence converges to a delta function, i.e., $p(X|\theta)$ is identifiable

$$p(X|T_i) \rightarrow p(X)$$

usually even for unidentifiable $p(X|\theta)$ 

Bayes Estimator

Bayes estimator minimizes the average risk $z(\hat{\theta})$ for $c_{\hat{\theta},\theta} = (\hat{\theta} - \theta)^2$
loss function:

$$\hat{\theta} = E\{\theta | T_i\} = \int \theta p(\theta | T_i) d\theta$$

also symmetric and convex cost functions or pdf symmetric about mean

$$z(\hat{\theta}) = E_{\theta}\{E\{(\hat{\theta} - \theta)\}\} = \int \int \dots \int (\hat{\theta} - \theta) p(X_1, \dots, X_n, \theta) dX_1 \dots dX_n d\theta$$

average risk for an estimate $\hat{\theta}$

Bayes estimator from (1)



Nonparametric Distributions - Histogram Approach

$l = 1 \dots X$ quantized into k equal intervals

$$\hat{p}(X | \omega_i) = h(X | \omega_i) = \frac{n_{ij}}{n_i}$$

n_{ij} number of $X_k \in S_i$ in the j -th interval

$n_i = \text{card } S_i$

⊖ $l > 1 \Rightarrow N_{bins} = k^l \longrightarrow$ only small k, l



Nonparametric Distributions - Parzen Estimation

assumption:

X continuous, given $X_k \in S_i \longrightarrow p(X = (X_k \pm \Delta) | \omega_i) \neq 0$

$$\hat{p}(X | \omega_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} K(X, X_j)$$

kernel properties:

- ① $K(X, X_j) = \rho^{-l} f\left(\frac{\delta(X, X_j)}{\rho}\right)$
- ② parameter ρ satisfies $\lim_{n_i \rightarrow \infty} \rho^l(n_i) = 0$
- ③ $\delta(X, X_j)$ metric
- ④ $\int K(X, X_j) dX = 1$ for non-negative $f()$ (alternatively f mon. decreasing, $f_{max} \rightarrow \delta(X, X_j) = 0$)

⊕ 2) + 4) $\Rightarrow \hat{p}(X | \omega_i)$ unbiased, consistent estimate



Nonparametric Distributions - Parzen Estimation

assumption:

X continuous, given $X_k \in S_i \longrightarrow p(X = (X_k \pm \Delta) | \omega_i) \neq 0$

$$\hat{p}(X | \omega_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} K(X, X_j)$$

kernel properties:

- ① $K(X, X_j) = \rho^{-l} f\left(\frac{\delta(X, X_j)}{\rho}\right)$
- ② parameter ρ satisfies $\lim_{n_i \rightarrow \infty} \rho^l(n_i) = 0$
- ③ $\delta(X, X_j)$ metric
- ④ $\int K(X, X_j) dX = 1$ for non-negative $f()$ (alternatively f mon. decreasing, $f_{max} \rightarrow \delta(X, X_j) = 0$)

⊕ 2) + 4) $\Rightarrow \hat{p}(X | \omega_i)$ unbiased, consistent estimate



Parzen Kernels

Gaussian Kernel

$$K(X, X_j) = \left[(\rho^2 2\pi)^l |Q| \right]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\rho^2} (X - X_j)^T Q (X - X_j) \right\}$$

Q arbitrary, better $Q = \hat{\Sigma}_i$

Hyperspheric Kernel

$$K(X, X_j) = \begin{cases} v^{-1} & \text{if } \delta(X, X_j) \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

$\delta(X, X_j)$ Euclidean distance, v volume of a hypersphere with radius ρ



Parzen Kernels

Gaussian Kernel

$$K(X, X_j) = \left[(\rho^2 2\pi)^l |Q| \right]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\rho^2} (X - X_j)^T Q (X - X_j) \right\}$$

Q arbitrary, better $Q = \hat{\Sigma}_i$

Hyperspheric Kernel

$$K(X, X_j) = \begin{cases} v^{-1} & \text{if } \delta(X, X_j) \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

$\delta(X, X_j)$ Euclidean distance, v volume of a hypersphere with radius ρ



Parzen Kernels

Hypercubic Kernel

$$K(X, X_j) = \begin{cases} (2\rho)^{-l} & \text{if } \delta(X, X_j) \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

$\delta(X, X_j) = \max_k |X_{.k} - X_{j.k}|$ Chebyshev distance

a value satisfying kernel condition (2):

$$\rho(n_i) = n_i^{-\frac{\eta}{l}}$$

$\eta \in (0; 1)$



Unsupervised Learning

- 1 unknown form of $p(X|\omega_i)$
 - mode separation methods
 - clustering
- ⊖ $\mathcal{X}$ partition, no pdf
- 2 mixture decomposition techniques
 $p(X|\omega_i)$ are of a known form



Unsupervised Learning

- 1 unknown form of $p(X|\omega_i)$
 - mode separation methods
 - clustering
- ⊖ $\mathcal{X}$ partition, no pdf
- 2 mixture decomposition techniques
 $p(X|\omega_i)$ are of a known form



UNISEP

projection method

- 1 Compute the eigenvector u associated with the max. eigenvalue of the sample covariance matrix and project the data on u , i.e., $\tilde{x}_i = u_i^T X$.
- 2 Obtain a histogram approximation of the marginal pdf $p(\tilde{X})$.
- 3 Determine the points of local minima of $p(\tilde{X})$ between all successive modes of the pdf and partition the data by hyperplanes perpendicular to u and intersecting the eigenvector at these points.
- 4 If no such point is determined, repeat for the next eigenvector.
- 5 Repeat the procedure for each individual subset that has been obtained by partitioning the complete sample set.



UNISEP

projection method

- 1 Compute the eigenvector u associated with the max. eigenvalue of the sample covariance matrix and project the data on u , i.e., $\tilde{x}_i = u_i^T X$.
- 2 Obtain a histogram approximation of the marginal pdf $p(\tilde{X})$.
- 3 Determine the points of local minima of $p(\tilde{X})$ between all successive modes of the pdf and partition the data by hyperplanes perpendicular to u and intersecting the eigenvector at these points.
- 4 If no such point is determined, repeat for the next eigenvector.
- 5 Repeat the procedure for each individual subset that has been obtained by partitioning the complete sample set.

- ⊖ fails on linearly unseparable sets
- ⊖ impossible to establish whether a peak corresponds to the actual data mode or to the noise
- ⊖ meaningless for a small data set



Unsupervised Bayesian Parametric Learning

assumptions

- K is known
- class priors $P(\omega_i) \quad \forall i = 1, \dots, K$ are known
- the forms for $p(X|\omega_i, \theta_i)$ are known except for some unknown parameters θ_i
- prior pdf $p(\theta_1, \dots, \theta_K)$ is known (Bayesian learning)
- the rest of knowledge about $\theta_i \quad \forall i$ is available from samples from the mixture density

$$p(X | \theta_1, \dots, \theta_K) = \sum_{i=1}^K p(X | \omega_i, \theta_i) P(\omega_i)$$

problem - using samples from $p(X | \theta_1, \dots, \theta_K)$ to estimate unknown parameters $\theta = [\theta_1, \dots, \theta_K]$



Unidentifiable Examples

continuous pdf (GMM)

$$p(X|\theta) = \frac{P(\omega_1)}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(X - \theta_1)^2\right\} + \frac{P(\omega_2)}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(X - \theta_2)^2\right\}$$

if $P(\omega_1) = P(\omega_2) \Rightarrow \theta_1, \theta_2$ can be interchanged without affecting $p(X|\theta) \rightarrow$ no identifiable solution



Unidentifiable Examples

discrete binary Bernoulli mixture

$$p(X|\theta) = \frac{1}{2}\theta_1^X(1 - \theta_1)^{1-X} + \frac{1}{2}\theta_2^X(1 - \theta_2)^{1-X}$$

$$p(X|\theta) = \begin{cases} \frac{1}{2}(\theta_1 + \theta_2) & \text{if } X = 1 \\ 1 - \frac{1}{2}(\theta_1 + \theta_2) & \text{if } X = 0 \end{cases}$$

e.g., $p(X = 1|\theta) = 0.6 \rightarrow p(X = 0|\theta) = 0.4$ $p(X|\theta)$ is known but we cannot determine θ and so component distributions the only knowledge is $\theta_1 + \theta_2 = 1.2$

⊕ identifiable mixtures - Poisson, gamma, Cauchy, nD Gaussian, nD exponential

⊖ unidentifiable mixtures - uniform, discrete, Bernoulli



Unidentifiable Examples

discrete binary Bernoulli mixture

$$p(X|\theta) = \frac{1}{2}\theta_1^X(1 - \theta_1)^{1-X} + \frac{1}{2}\theta_2^X(1 - \theta_2)^{1-X}$$

$$p(X|\theta) = \begin{cases} \frac{1}{2}(\theta_1 + \theta_2) & \text{if } X = 1 \\ 1 - \frac{1}{2}(\theta_1 + \theta_2) & \text{if } X = 0 \end{cases}$$

e.g., $p(X = 1|\theta) = 0.6 \rightarrow p(X = 0|\theta) = 0.4$ $p(X|\theta)$ is known but we cannot determine θ and so component distributions the only knowledge is $\theta_1 + \theta_2 = 1.2$

⊕ identifiable mixtures - Poisson, gamma, Cauchy, nD Gaussian, nD exponential

⊖ unidentifiable mixtures - uniform, discrete, Bernoulli



Unsupervised Parametric Bayesian Learning - Complexity

$${}^tT = {}^{t-1}T + \{X_t\}$$

$$\theta^T = [\theta_1^T, \dots, \theta_K^T, P_1, \dots, P_K]$$

assumptions

1

$$p(X_t|\theta) = \sum_{k=1}^K P_k p_k(X_t|\theta_k)$$

2 conditional independence

$$p({}^tT|\theta) = \prod_{i=1}^t p(X_i|\theta) = \prod_{i=1}^t \sum_{j=1}^K P_j p_j(X_i|\theta_j)$$



Unsupervised Parametric Bayesian Learning 2

then

$$\begin{aligned} p(\theta | {}^t T) &= \frac{p(X_t | \theta) p(\theta | {}^{t-1} T)}{\int p(X_t | \theta) p(\theta | {}^{t-1} T) d\theta} \\ &= \frac{\sum_{k=1}^K P_k p_k(X_t | \theta_k) p(\theta | {}^{t-1} T)}{\int \sum_{k=1}^K P_k p_k(X_t | \theta_k) p(\theta | {}^{t-1} T) d\theta} \end{aligned}$$

- ⊖ $p(\theta_k | T)$ depends on distribution of other classes
- ⊖ $p(T | \theta) \neq g(\theta, \Xi) h(T)$ due to 1. → no reproducing density for UL
→ $p(\theta | T)$ numerical method



Gaussian Mixture Example - Moment Method

$$f(X | \theta) = p g_1(X | \theta_1) + (1 - p) g_2(X | \theta_2)$$

$$g_1(\cdot) \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad g_2(\cdot) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$E\{X\} = p \mu_1 + (1 - p) \mu_2$$

$$E\{X^2\} = p(\mu_1^2 + \sigma_1^2) + (1 - p)(\mu_2^2 + \sigma_2^2)$$

$$E\{X^3\} = p(\mu_1^3 + 3\mu_1\sigma_1^2) + (1 - p)(\mu_2^3 + 3\mu_2\sigma_2^2)$$

if $\sigma_1^2 = \sigma_2^2$ known and $\mu_2 = 0$ then:

$$E\{X\} = p \mu_1$$

$$E\{X^2\} = p(\mu_1^2 + \sigma^2)$$

$$E\{X^3\} = p(\mu_1^3 + 3\mu_1\sigma^2)$$

$$\Rightarrow \mu_1 = 0 \quad p = \frac{2}{3}$$

⊖ computationally complex for $K \nearrow \vee d \nearrow$



Unsupervised ML Learning

data in T from the mixture pdf, θ fixed but unknown

$$p(X|\theta) = \sum_{i=1}^K p(X|\omega_i, \theta_i) P(\omega_i)$$

likelihood of T

$$p(T|\theta) = \prod_{j=1}^n p(X_j|\theta)$$

$$\hat{\theta} = \arg \max_{\theta} \{p(T|\theta)\}$$

log-likelihood

$$L(\theta) = \log p(T|\theta) = \sum_{j=1}^n \log p(X_j|\theta)$$



Unsupervised ML Learning 2

$$\nabla_{\theta_i} L = \sum_{j=1}^n \frac{1}{p(X_j|\theta)} \nabla_{\theta_i} \left[\sum_{i=1}^K p(X_j|\omega_i, \theta_i) P(\omega_i) \right]$$

assuming functionally independent θ_i, θ_j

$$\nabla_{\theta_i} L = \sum_{j=1}^n P(\omega_i|X_j, \theta_i) \nabla_{\theta_i} \log p(X_j|\omega_i, \theta_i)$$

the maximum likelihood estimate

$$\nabla_{\theta_i} L = 0$$



EM Algorithm - (Expectation & Maximization)

⊕ iterative ML estimation from incomplete observations

$X \in \mathcal{X}$ measurements (incomplete data)
 $\omega \in \Omega$ hidden, unobservable (complete) data
 $\omega \rightarrow X(\omega)$ $\omega \in \Omega(X) \subset \Omega$

$$p(X|\theta) = \int_{\Omega(X)} p(\omega|\theta) d\omega$$

E-step: $f(\theta^k|\theta) = E\{\log p(\omega|\theta) | X, \theta^k\} = \int_{\omega} \log p(\omega|\theta) p(\omega|X, \theta^k) d\omega$

M-step: $\theta^{k+1} = \arg \max_{\theta} \{f(\theta^k|\theta)\}$



EM Algorithm - Example

$K = 4, n = 197, X = [125, 18, 20, 34]$ multinomial
 $p_1 = \frac{2+\alpha}{4}, p_2 = \frac{1-\alpha}{4}, p_3 = \frac{1-\alpha}{4}, p_4 = \frac{\alpha}{4} \quad 0 \leq \alpha \leq 1 \quad \sum p_i = 1$

$$P(X) = \frac{(\sum_{i=1}^4 x_i)!}{\prod_{i=1}^4 x_i!} \prod_{i=1}^4 p_i^{x_i}$$

$\omega = [\omega_1, \dots, \omega_5], x_1 = \omega_1 + \omega_2, x_2 = \omega_3, x_3 = \omega_4, x_4 = \omega_5$
 ω_1, ω_2 unknown, $\omega_3, \dots, \omega_5$ known, $P(\omega)$ multinomial

E-step:

$$E\{\omega_1^k\} = 125 \frac{\frac{1}{2}}{\frac{2+\alpha^k}{4}} \quad E\{\omega_2^k\} = 125 \frac{\frac{\alpha^k}{4}}{\frac{2+\alpha^k}{4}}$$

M-step:

$$\frac{\partial P(\omega, \alpha)}{\partial \alpha} = 0 \rightarrow \alpha^{k+1} = \frac{\omega_2^k + 34}{\omega_2^k + 34 + 18 + 20}$$

$$\alpha^8 = 0.626821484 \quad \alpha^* = 0.626821498$$



EM Properties

$$p(X_r | \theta_j) \leftarrow \begin{cases} \text{discrete} \\ \text{continuous} \end{cases} \quad \text{parametric pdf}$$

$$P(\omega_j) \geq 0 \quad \forall j, \quad \sum_{j=1}^K P(\omega_j) = 1$$

$$p(X_r | \theta_1, \dots, \theta_K) = \sum_{j=1}^K P(\omega_j) p(X_r | \theta_j)$$

$$\mathcal{L} = \sum_{X_r \in \mathcal{X}} \log \left[\sum_{j=1}^K P(X_r | \theta_j) P(\omega_j) \right]$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \quad \text{no explicit solution}$$



EM Properties

T: $\mathcal{L}^{t+1} - \mathcal{L}^t \geq 0 \quad t = 0, 1, \dots$

T: if $\lim_{t \rightarrow \infty} \mathcal{L} = \mathcal{L}^* < \infty \Rightarrow$

- $\mathcal{L}^*$ local maximum
- $\lim_{t \rightarrow \infty} \|P_{\omega}^{t+1} - P_{\omega}^t\| = 0$
- $\lim_{t \rightarrow \infty} \|P_{\omega|X}^{t+1} - P_{\omega|X}^t\| = 0$

$$P_{\omega}^t = [P^t(\omega_1), \dots, P^t(\omega_K)]^T$$

$$P_{\omega|X}^t = [P^t(\omega_j | X_r)]_{j=1}^K, \quad \forall r$$



E-step:

$$P^{t+1}(\omega_j | X_r) = \frac{P^{t+1}(\omega_j) P(X_r | \theta_j^{t+1})}{P^{t+1}(X_r | \theta_1, \dots, \theta_K)}$$

M-step:

$$P^{t+1}(\omega_j) = \frac{1}{|\mathcal{X}|} \sum_{\forall X_r \in \mathcal{X}} P^t(\omega_j | X_r)$$

- continuous + discrete

$$\theta_j^{t+1} = \arg \max_{\theta_j \in \Theta} \left\{ \sum_{\forall X_r \in \mathcal{X}} P^t(\omega_j | X_r) \log P(X_r | \theta_j) \right\}$$

- discrete unknown form

$$P^{t+1}(X_r = \zeta | \theta_i) = \frac{1}{|\mathcal{X}| P^{t+1}(\omega_i)} \sum_{\forall X_s} \delta(X_s, \zeta) P^t(\omega_i | X_s)$$



$$p(X_1, \dots, X_N) = \sum_{j=1}^K P(\omega_j) p(X_1, \dots, X_N | \theta_j)$$

assumption

$$p(X_1, \dots, X_N | \theta_j) = \prod_{X_r} p(X_r | \theta_j)$$

conditional independence not restrictive for discrete mixtures
($K \nearrow$)

$$\mathcal{L} = \frac{1}{|\mathcal{X}|} \sum_{\mathcal{X}} \log \left[\sum_{j=1}^K P(\omega_j) \prod_{X_r \in \mathcal{X}} P(X_r | \theta_j) \right]$$



E-step:

$$P^{t+1}(\omega_j | X_1, \dots, X_N) = \frac{P^{t+1}(\omega_j) P(X_1, \dots, X_N | \theta_j^{t+1})}{\sum_{i=1}^K P^{t+1}(X_1, \dots, X_N | \theta_i) P^{t+1}(\omega_i)}$$

M-step:

$$P^{t+1}(\omega_j) = \frac{1}{|\mathcal{X}|} \sum_{\forall X \in \mathcal{X}} P^t(\omega_j | X_1, \dots, X_N)$$

$$P^{t+1}(X_r = \zeta | \theta_i) = \frac{1}{|\mathcal{X}| P^{t+1}(\omega_i)} \sum_{\forall X_s} \delta(X_s, \zeta) P^t(\omega_i | X_1, \dots, X_N)$$



$$p(X_n | X_C) = \sum_{i=1}^K P(\omega_{i,C}) p(X_n | \theta_i) \quad X_C \subset \mathcal{X}$$

$$P(\omega_{i,C}) = \frac{P(\omega_i) \prod_{j \in C} p(X_j | \theta_i)}{\sum_{k=1}^K P(\omega_k) \prod_{j \in C} p(X_j | \theta_k)}$$

⊕ simple derivation of any marginal distribution



EM - Gaussian Mixture

$$p(X) = \sum_{j=1}^K P(\omega_j) \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu_j)^T \Sigma_j^{-1} (X - \mu_j) \right\}$$

$$\mathcal{L} = \frac{1}{|\mathcal{X}|} \sum_{\mathcal{X}} \log \left[\sum_{j=1}^K P(\omega_j) \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu_j)^T \Sigma_j^{-1} (X - \mu_j) \right\} \right]$$



EM - Gaussian Mixture 2

E-step:

$$P^{t+1}(\omega_j | X) = \frac{P^{t+1}(\omega_j) \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu_j)^T \Sigma_j^{-1} (X - \mu_j) \right\}}{\sum_{i=1}^K P^{t+1}(\omega_i) \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) \right\}}$$

M-step:

$$P^{t+1}(\omega_j) = \frac{1}{|\mathcal{X}|} \sum_{\forall X \in \mathcal{X}} P^t(\omega_j | X)$$

$$\mu_j^{t+1} = \frac{1}{\sum_{\forall X \in \mathcal{X}} P^t(\omega_j | X)} \sum_{\forall X \in \mathcal{X}} X P^t(\omega_j | X)$$

$$\Sigma_j^{t+1} = \frac{1}{\sum_{\forall X \in \mathcal{X}} P^t(\omega_j | X)} \sum_{\forall X \in \mathcal{X}} X X^T P^t(\omega_j | X) - \mu_j^{t+1} \mu_j^{t+1,T}$$



EM - Gaussian Mixture 3

- ◉ Σ_j^{t+1} possibly singular $\rightarrow$ new initialization
- ◉ $\Sigma_j = \Sigma \quad j = 1, \dots, K$

$$\Sigma_j^{t+1} = \frac{1}{|\mathcal{X}|} \sum_{\forall X \in \mathcal{X}} X X^T - \sum_{j=1}^K P^{t+1}(\omega_j) \mu_j^{t+1} \mu_j^{t+1,T}$$



EM Applications

- recognition (clustering) $P^{t+1}(\omega_j) \rightarrow$ prior class probability
assign X_r into ω_j if

$$\omega_j = \arg \max_{\omega_j} \{ P^{t+1}(\omega_i | X_r) \}$$

$$P^{t+1}(\omega_j | X_r) = \frac{P^{t+1}(\omega_j) P(X_r | \theta_j^{t+1})}{P^{t+1}(X_r | \theta_1, \dots, \theta_K)}$$

- unsupervised learning $P^{t+1}(\omega_j) \rightarrow$ probabilistic coefficient

$$\theta_j^{t+1} = \arg \max_{\theta_j \in \Theta} \left\{ \sum_{\forall X_r \in \mathcal{X}} P^t(\omega_j | X_r) \log P(X_r | \theta_j) \right\}$$

- data recovering - sampling from the estimated mixture
- texture synthesis



Supervised Versus Unsupervised Learning

SL not identifiable $p(X|\theta) \Rightarrow {}^1\hat{\theta}, {}^2\hat{\theta}, \dots$, the same $p(X)$

UL not identifiable $p(X|\theta) \Rightarrow {}^1\hat{\theta}, {}^2\hat{\theta}, \dots$, impossible to decompose

$p(X|\theta)$ i.e. $p(X|\omega_i, T) \not\rightarrow p(X|\omega_i)$

UL no simple analytical solution for $p(\theta|T)$

SL new sample $p(\theta|^t T) \sim p({}^t X|\omega_i, \theta_i)p(\theta|{}^{t-1} T)$

UL new sample - less effective learning

$p(\theta|^t T) \sim \sum_{j=1}^K p({}^t X|\omega_j, \theta_j)P(\theta_j)p(\theta|{}^{t-1} T)$



Training Set Selection

ideally $P(\text{error}) \sim f(T)$, T quality determines $P(\text{error})$

Required properties:

- representative (e.g., **noisy**, **object's changes**)
- balanced $|T_i| \approx |T_j| \forall j \neq i$ (e.g., **skin lesions** - 80% benign, 10% malign, 10% unknown, **fraud detection**, **natural disaster prediction**)
- T_i homogeneity - cluster analysis, statistical tests (chi-square, Kolmogorov-Smirnov)
- $\text{card}\{T\} \rightarrow \infty$ - too expensive
 $\hat{\Sigma}_i^{-1} \rightarrow \text{card}\{T_i\} \geq l+1$ better $\text{card}\{T_i\} \geq 10$
 minimal dispersion of $\hat{\Sigma}$ criterion ($P(\text{error})$?)
 $\hat{\Sigma}$ - Wishart distribution, $\rightarrow \text{card}\{T_i\} \geq 5$



Training Set Selection

ideally $P(\text{error}) \sim f(T)$, T quality determines $P(\text{error})$

Required properties:

- representative (e.g., **noisy**, **object's changes**)
- balanced $|T_i| \approx |T_j| \forall j \neq i$ (e.g., **skin lesions** - 80% benign, 10% malign, 10% unknown, **fraud detection**, **natural disaster prediction**)
- undersampling** delete $i X \in T_i$ if $|T_i| \gg |T_j|$ **information loss**
- oversampling** copies of existing samples **overfitting**
- synthesis** new samples synthesis **reliable model**
- T_i homogeneity - cluster analysis, statistical tests (chi-square, Kolmogorov-Smirnov)
- $\text{card}\{T\} \rightarrow \infty$ - too expensive
 $\hat{\Sigma}_i^{-1} \rightarrow \text{card}\{T_i\} \geq l+1$ better $\text{card}\{T_i\} \geq 10$
 minimal dispersion of $\hat{\Sigma}$ criterion ($P(\text{error})$?)
 $\hat{\Sigma}$ - Wishart distribution, $\rightarrow \text{card}\{T_i\} \geq 5$



Training Set Selection

ideally $P(\text{error}) \sim f(T)$, T quality determines $P(\text{error})$

Required properties:

- representative (e.g., **noisy**, **object's changes**)
- balanced $|T_i| \approx |T_j| \forall j \neq i$ (e.g., **skin lesions** - 80% benign, 10% malign, 10% unknown, **fraud detection**, **natural disaster prediction**)
- T_i homogeneity - cluster analysis, statistical tests (chi-square, Kolmogorov-Smirnov)
- $\text{card}\{T\} \rightarrow \infty$ - too expensive
 $\hat{\Sigma}_i^{-1} \rightarrow \text{card}\{T_i\} \geq l+1$ better $\text{card}\{T_i\} \geq 10$
 minimal dispersion of $\hat{\Sigma}$ criterion ($P(\text{error})$?)
 $\hat{\Sigma}$ - Wishart distribution, $\rightarrow \text{card}\{T_i\} \geq 5$



Training Set Selection

ideally $P(\text{error}) \sim f(T)$, T quality determines $P(\text{error})$

Required properties:

- representative (e.g., **noisy**, **object's changes**)
- balanced $|T_i| \approx |T_j| \forall j \neq i$ (e.g., **skin lesions** - 80% benign, 10% malign, 10% unknown, **fraud detection**, **natural disaster prediction**)
- T_i homogeneity - cluster analysis, statistical tests (chi-square, Kolmogorov-Smirnov)
- $\text{card}\{T\} \rightarrow \infty$ - too expensive
 $\hat{\Sigma}_i^{-1} \rightarrow \text{card}\{T_i\} \geq l + 1$ better $\text{card}\{T_i\} \geq 10l$
 minimal dispersion of $\hat{\Sigma}$ criterion ($P(\text{error})$?)
 $\hat{\Sigma}$ - Wishart distribution, $\rightarrow \text{card}\{T_i\} \geq 5l$



Training Set Selection 2

- if $\text{card}\{T_i\}$ insufficient
feature selection
 • $\Rightarrow I \searrow$



Training Set Selection 2

- if $\text{card}\{T_i\}$ insufficient
regularization
 • common $\hat{\Sigma}$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^K n_i \hat{\Sigma}_i$$

- $\hat{\Sigma}_i(\lambda) = \lambda \hat{\Sigma} + (1 - \lambda) \hat{\Sigma}_i$ $\lambda \in \langle 0; 1 \rangle$
- shrinking $c \in \langle 0; 1 \rangle$

$$\hat{\Sigma}_i(c) = (1 - c) \hat{\Sigma}_i + \frac{c}{n} \text{tr}(\hat{\Sigma}_i) \mathbf{I}$$



Training Set Selection 2

- if $\text{card}\{T_i\}$ insufficient
pooled sample covariance

$$\hat{\Sigma}_p = \frac{\sum_{i=1}^K (n_i - 1) \hat{\Sigma}_i}{N - K}$$

$$N = \sum_{i=1}^K n_i$$

$$n_i = \text{card}\{T_i\}$$

$\hat{\Sigma}_p$ consistent iff $\hat{\Sigma}_i = \hat{\Sigma}_j \forall i \neq j$



Training Set Selection 2

- if $\text{card}\{T_i\}$ insufficient

LOOC estimate

$$\hat{\Sigma}_i^{\text{LOOC}} = \begin{cases} (1 - \alpha_i)\text{diag}(\Sigma_i) + \alpha_i\Sigma_i, & 0 \leq \alpha_i \leq 1 \\ (2 - \alpha_i)\Sigma_i + (\alpha_i - 1)\Sigma, & 1 < \alpha_i \leq 2 \\ (3 - \alpha_i)\Sigma + (\alpha_i - 2)\text{diag}(\Sigma), & 2 < \alpha_i \leq 3 \end{cases}$$

$$\Sigma = \frac{1}{K} \sum_{i=1}^K \hat{\Sigma}_i$$

optimization Gaussian log ML over α_i



Training Set Selection 3

- Gaussian assumption - seldom true, small influence to $P(\text{error})$
- class selection
 - separable classes
 - classes with information values
- feature vector dimensionality

$$I \rightarrow \infty \quad P(\text{error}) \rightarrow 0$$

the Hughes phenomenon - for $\text{card}\{T_i\} < \infty$
 $P(\text{error}) \searrow \nearrow$, i.e. $\exists! I^*$

- ⊖ biased $\hat{\mu}, \hat{\Sigma}$
- ⊖ a class is composed from subclasses
- ⊖ change in measurement conditions (atmospheric changes)



Training Set Selection 3

- Gaussian assumption - seldom true, small influence to $P(\text{error})$
- class selection
 - separable classes
 - classes with information values
- feature vector dimensionality

$$I \rightarrow \infty \quad P(\text{error}) \rightarrow 0$$

the Hughes phenomenon - for $\text{card}\{T_i\} < \infty$
 $P(\text{error}) \searrow \nearrow$, i.e. $\exists! I^*$

- ⊖ biased $\hat{\mu}, \hat{\Sigma}$
- ⊖ a class is composed from subclasses
- ⊖ change in measurement conditions (atmospheric changes)



Training Set Selection 3

- Gaussian assumption - seldom true, small influence to $P(\text{error})$
- class selection
 - separable classes
 - classes with information values
- feature vector dimensionality

$$I \rightarrow \infty \quad P(\text{error}) \rightarrow 0$$

the Hughes phenomenon - for $\text{card}\{T_i\} < \infty$
 $P(\text{error}) \searrow \nearrow$, i.e. $\exists! I^*$

- ⊖ biased $\hat{\mu}, \hat{\Sigma}$
- ⊖ a class is composed from subclasses
- ⊖ change in measurement conditions (atmospheric changes)

