Lecture 5

STAT261: Introduction to Machine Learning

Lecture 5, April 11

Outline: April 11

- Lectures 5
- Probabilistic view
  - Uncertainty in evidence and predictors
- Bayes Estimation: Overview and Review
  - Priors, Bayes Loss Functions
- Parameter Learning and Maximum Likelihood
- Bayesian MAP Parametric Estimation
- Bias Variance Trade-off
- Parametric Classification
  - Maximum Likelihood Classification
  - MAP Classification

Statistical Learning:

- Given data $(x_i, y_i), i = 1, \ldots, n$
- $x_i \in R^p$ vectors of covariates or predictors
  - Also called independent variables
- Supervised: $y_i \in R$ target or dependent variable
- Want to learn the function:

$$y \approx f(x)$$

- Why?
  - Prediction: Given past data, predict response on future samples
  - Inference: Functions indicates relation between variables

Bayes Estimation: Choices

- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta\, p(\mathcal{X}|\theta)$
- Treat θ as a random var with prior $p(\theta)$ $p(\theta)$
- Bayes' rule: $p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)\, p(\theta) \,/\, p(\mathcal{X})$

  -Full $p(\mathcal{X}) = \int p(x|\theta)\, p(\theta)\, d\theta$
- Maximum a Posteriori (MAP): $\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta\, p(\theta|\mathrm{X})$
- Bayes': $\theta_{\mathrm{Bayes'}} = \mathrm{E}[\theta|\mathrm{X}] = \int \theta\, p(\theta|\mathrm{X})\, d\theta$

Pros and Cons of Bayesian Approach

- Pros:
  - Can exploit prior knowledge
  - Will generally improve estimation error
    - If assumptions are correct
- Cons:
  - Relies on our assumptions being correct
    - If incorrect, estimates can be very wrong
  - Estimates are biased by the assumptions
  - Requires a precise specification of the prior

How Do We get a Prior?

- Looks at past examples and fit a probability distribution
  - We did this in the last few lectures
  - Requires: We have sufficient number of samples
  - Assumption that the future will be like the past
- Use expert knowledge or physical modeling
  - Reliable, but many systems are too complex to model
- Many approaches use a combination

Loss Function

- Consider estimator $\hat{\theta} = g(x)$
- What is a good estimator?
- Suppose we have a loss function or risk: $L(\theta, \hat{\theta})$
  - Represents the cost of selecting $\hat{\theta}$ when true value is θ
- Bayes risk minimization: Given x,

$$\hat{\theta} = \arg\min_{\hat{\theta}} E\left[L(\theta, \hat{\theta})|x\right]$$
$$= \arg\min_{\hat{\theta}} \int L(\theta, \hat{\theta}) p(\theta|x)\, d\theta$$

MMSE and MAP

- MMSE minimizes squared loss:

$$L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$$

- MAP: Suppose θ is discrete

$$L(\theta,\hat{\theta}) = \begin{cases} 1 & \text{if } \theta \neq \hat{\theta} \\ 0 & \text{if } \theta = \hat{\theta} \end{cases}$$

  - $E(L(\theta,\hat{\theta})|x) = P(\theta \neq \hat{\theta}|x) =$ probability of error
  - MAP minimizes probability of error

Outline: April 11

- Lectures 5
- Probabilistic view
  - Uncertainty in evidence and predictors
- Bayes Estimation: Overview and Review
  - Priors, Bayes Loss Functions
- Parameter Learning and Maximum Likelihood
- Bayesian MAP Parametric Estimation
- Bias Variance Trade-off
- Parametric Classification
  - Maximum Likelihood Classification
  - MAP Classification

Parametric Learning

- $\mathcal{X} = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid \theta)$ and estimate θ, using X

  e.g., N (μ, σ^2) where $\theta = \{ \mu, \sigma^2 \}$
- Or assume $y = f(x)$ has some parametric form, $f(x, \beta)$
- Examples:
  - Linear model: $f(x,\beta) = \beta_0 + \beta_1 x$
  - Sinusoid with unknown frequency:

    $$f(x,\beta) = \beta_0 \cos(\beta_1 x + \beta_2)$$
  - Exponential: $f(x,\beta) = \beta_0 e^{-\beta_1 x}$
- Many possibilities
- Problem: Learn the parameter vector β from data.

Maximum Likelihood Parametric Estimation

- Likelihood of θ given the sample $\mathcal{X}$

$$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$$

- Log likelihood

$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

- Maximum likelihood estimator (MLE)

$$\theta^* = \operatorname{argmax}_\theta \mathcal{L}(\theta|\mathcal{X})$$

Example: MLE of an Exponential

- Data: $\boldsymbol{x} = (x_1, \dots, x_n)$ i.i.d. $p(x_i|\lambda) = \frac{1}{\lambda} e^{-x_i/\lambda}$
- MLE: $\hat{\lambda} = \arg\max_{\lambda} p(\boldsymbol{x}|\lambda) = \arg\max_{\lambda} \mathcal{L}(\lambda)$
- Log likelihood:

$$\mathcal{L}(\lambda) \coloneqq \ln p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{n} \ln p(x_i|\lambda) = -n \ln \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} x_i$$

- Take derivative:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = 0 \Rightarrow \frac{n}{\lambda} = \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Conclusion: MLE for an exponential is the sample mean

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, x in {0,1}

$$P(x) = p_o^{x}(1-p_o)^{(1-x)}$$

$$\mathcal{L}(p_o|\mathcal{X}) = \log \prod_t p_o^{x^t}(1-p_o)^{(1-x^t)}$$

MLE: $p_o = \sum_t x^t / N$

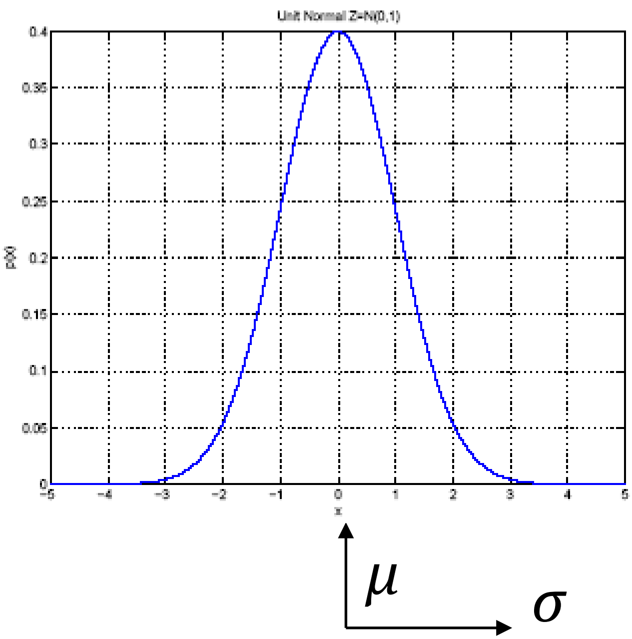
- Multinomial: K>2 states, x_i in {0,1}

$$P(x_1,x_2,\ldots,x_K) = \prod_i p_i^{x_i}$$

$$\mathcal{L}(p_1,p_2,\ldots,p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$$

MLE: $p_i = \sum_t x_i^t / N$

MLE: Gaussian Parameter Estimation



- $p(x) = \mathcal{N}(\mu, \sigma^2)$

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

- MLE for μ and σ^2:

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t \left(x^t - m\right)^2}{N}$$

Bayes MAP: Gaussian Parameter Estimation

- Data $x^t \sim N(\theta, \sigma_o^2)$ and prior $\theta \sim N(\mu, \sigma^2)$
- Find MAP estimate of the mean
- Maximum a Posteriori (MAP):

$$
\begin{aligned}
\theta_{\text{MAP}} &= \operatorname{argmax}_\theta\, p(\theta \mid \mathrm{X}) \\
&= \operatorname{argmax}_\theta\, p(X \mid \theta)\, p(\theta) / p(\mathrm{X}) \\
&= \operatorname{argmax}_\theta\, p(X \mid \theta)\, p(\theta) \\
&= \operatorname{argmax}_\theta\, (\ln p(X|\theta) + \ln p(\theta))
\end{aligned}
$$

Bayes MAP: Gaussian Parameter (Contd)

- θ_{MAP} = argmax of the log posterior
- The log posterior is

$$\ln p(\theta|X) = \ln p(X|\theta) + \ln p(\theta)$$
$$= -\sum_{t=1}^{n} \frac{(x^t - \theta)^2}{2\sigma_0^2} - \frac{(\theta - \mu)^2}{2\sigma^2} + const$$

- Quadratic in $\theta \Rightarrow p(\theta|X)$ is Gaussian
- Thus, $\theta_{MAP} = \theta_{Bayes}$

$$= \frac{N\sigma^2}{N\sigma^2 + \sigma_0^2} m + \frac{\sigma_0^2}{N\sigma^2 + \sigma_0^2} \mu$$

  - Linear combination of MLE m and prior μ
  - Note: $\theta_{\mathrm{ML}} = m$ = sample mean
    Now we have weighted sample mean and prior mean

Outline: April 11

- Lectures 5
- Probabilistic view
  - Uncertainty in evidence and predictors
- Bayes Estimation: Overview and Review
  - Priors, Bayes Loss Functions
- Parameter Learning and Maximum Likelihood
- Bayesian MAP Parametric Estimation
- Bias Variance Trade-off
- Parametric Classification
  - Maximum Likelihood Classification
  - MAP Classification

Estimators, Bias and Variance

- Estimator: Maps X to $\hat{\theta}$ = estimate of θ
  - $\hat{\theta} = \hat{\theta}(X)$ a function of the data
  - For any parameter, θ, $\hat{\theta}(X)$ is a random variable
- Bias: $\mathrm{Bias}(\hat{\theta}|\theta) = E(\hat{\theta} - \theta\,|\theta)$
- Variance $\mathrm{var}(\hat{\theta}|\theta) = E\left(\left(\hat{\theta} - E(\hat{\theta}|\theta)\right)^2 \middle|\theta\right)$
- Note: Bias and variance depend on true parameter θ

- Bias + variance formula: Mean squared error (MSE)

$$E\left((\hat{\theta} - \theta)^2\right) = \mathrm{Bias}^2(\hat{\theta}|\theta) + \mathrm{var}(\hat{\theta}|\theta)$$

Bias/Var Example: Gaussian Parameter Est

- Data $x^t \sim N(\theta, \sigma_0^2)$, prior: $\theta \sim N(\mu, \sigma^2)$
- MLE: $\hat{\theta} = \text{argmax}_\theta \, p(X \mid \theta)$

  $= \text{sample mean} = \text{m}$

  - Bias: $E(\hat{\theta}|\theta) - \theta = 0$

    Unbiased
  - Variance: $var(\hat{\theta}|\theta) = \sigma_0^2/N$

Example: MAP Gaussian Parameter Est

- Data $x^t \sim N(\theta, \sigma_0^2)$, prior: $\theta \sim N(\mu, \sigma^2)$
- MAP or Bayes estimator:
  - $\hat{\theta} = \alpha m + (1-\alpha)\mu,\ \alpha = N\sigma^2/(N\sigma^2 + \sigma_0^2)$
  - Weights prior μ with estimate from evidence m
  - Bias: $E(\hat{\theta}|\theta) - \theta = (1-\alpha)(\mu - \theta)$
  - Variance: $var(\hat{\theta}|\theta) = \alpha^2\sigma_0^2/N$
  - Variance is smaller than MLE,
    but bias grows as θ is different from μ

Vector Parameters

- Suppoose $\theta = (\theta_1, \dots, \theta_p)^T$ is a (column) vector.
- Bias is a vector: $\mathrm{Bias}(\hat{\theta}|\theta) = E(\hat{\theta} - \theta\,|\theta)$
- Variance is a matrix:

$$\mathrm{var}(\hat{\theta}|\theta) = E\left(\hat{\theta} - E(\hat{\theta}|\theta)\left(\hat{\theta} - E(\hat{\theta}|\theta)\right)^T \middle|\theta\right)$$

- Bias + Variance formula provides a matrix

Outline: April 11

- Lectures 5
- Probabilistic view
  - Uncertainty in evidence and predictors
- Bayes Estimation: Overview and Review
  - Priors, Bayes Loss Functions
- Parameter Learning and Maximum Likelihood
- MAP and MMSE Parametric Estimation
- Bias Variance Trade-off
- Parametric Classification

Classification: Maximum Likelihood

- Classification: Maximum Likelihood
- Input: $\boldsymbol{x} = [x_1, x_2]^T$,Output: C : {0,1}
- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 \mid x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 \mid x_1, x_2) > P(C = 0 \mid x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

MAP: Classification Bayes' Rule

posterior

prior

likelihood

$$P(C \mid \mathbf{x}) = \frac{P(C)\, p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

evidence

$$P(C=0) + P(C=1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} \mid C=1)P(C=1) + p(\mathbf{x} \mid C=0)P(C=0)$$

$$p(C=0 \mid \mathbf{x}) + P(C=1 \mid \mathbf{x}) = 1$$

MAP: Parametric Classification

Pick from maximal "map" estimate of class from observation

$$g_i(x) = p(x \mid C_i)P(C_i)$$

or

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Parametric Classification

- Given the sample

$$\mathcal{X} = \{x^t, r^t\}_{t=1}^{N}$$

$$x \in \Re$$

$$r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

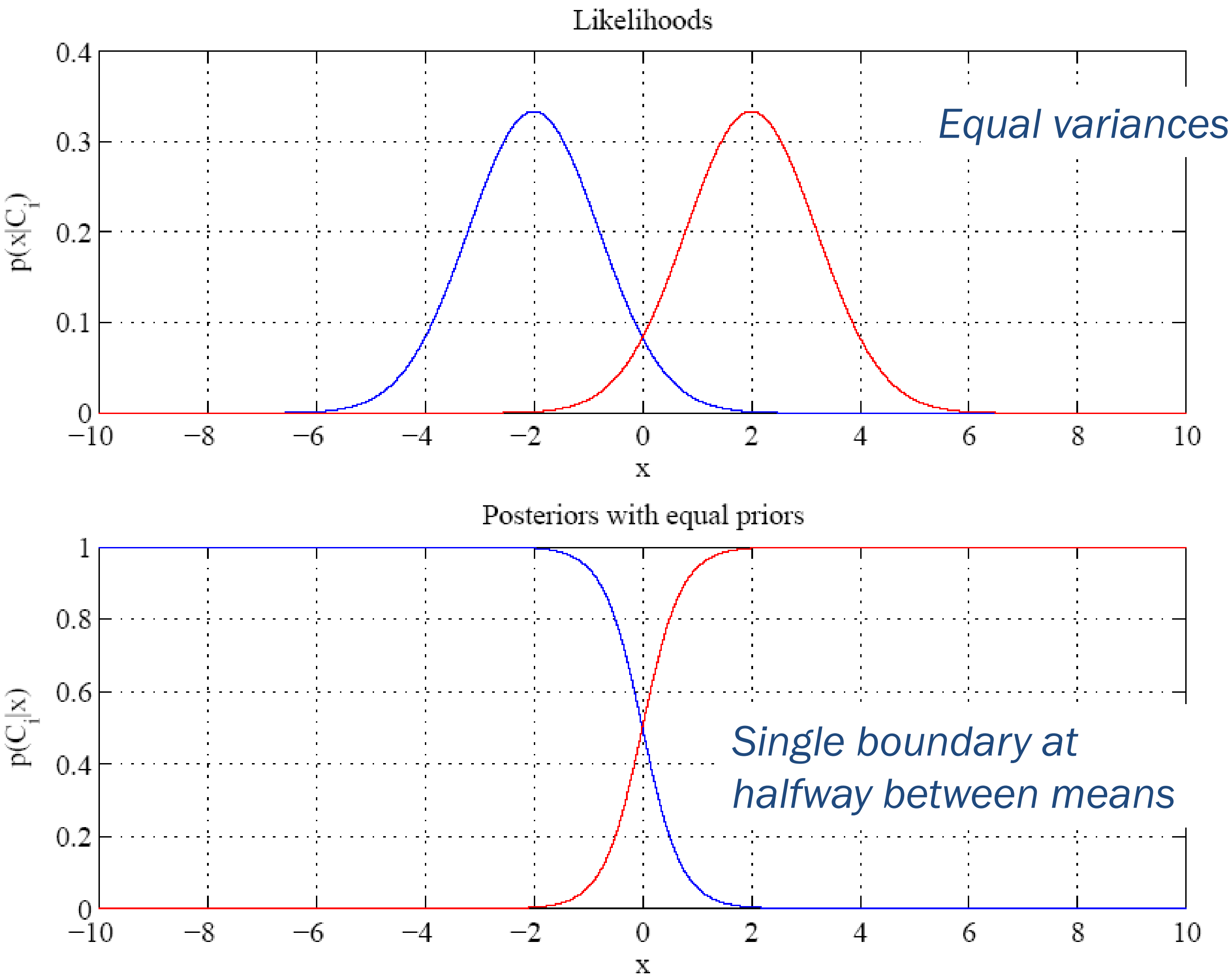
- ML estimates are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

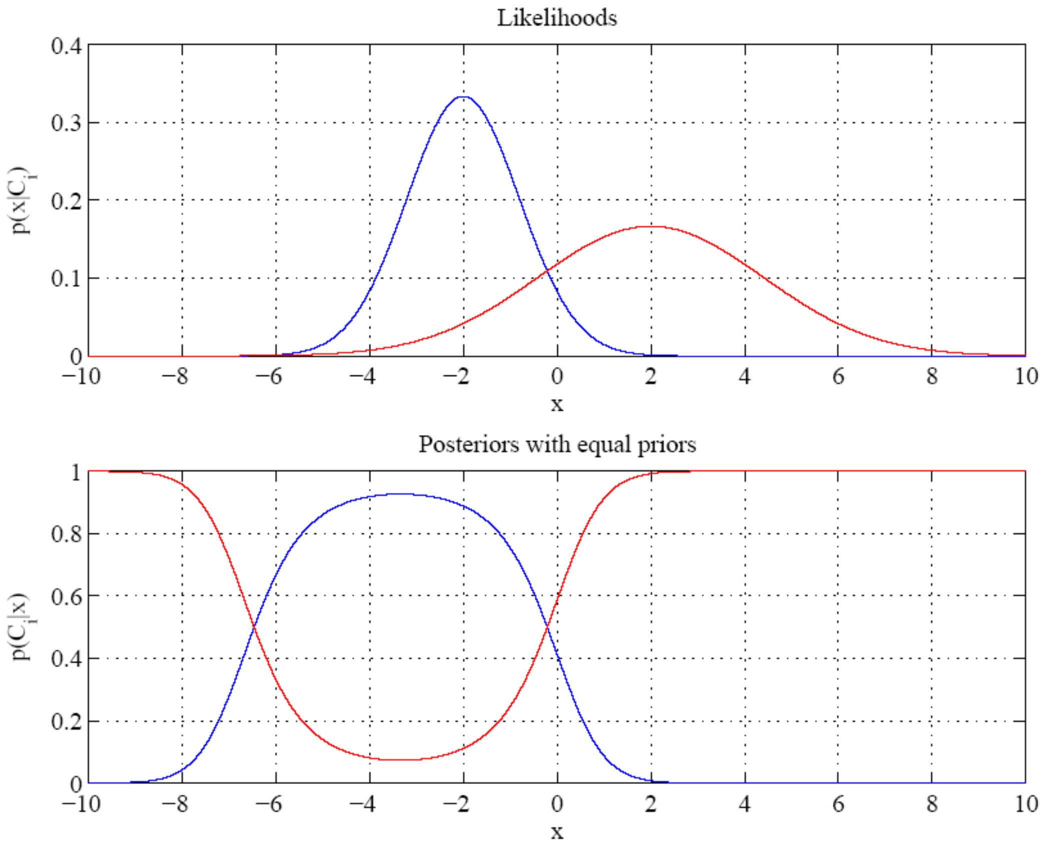
- Discriminant becomes

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{\left(x - m_i\right)^2}{2 s_i^2} + \log \hat{P}(C_i)$$

Parametric Classification



Parametric Classification



Variances are different

Two boundaries

Bayes' Rule: $K>2$ Classes

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)P(C_i)}{p(\mathbf{x})}$$

$$= \frac{p(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k)P(C_k)}$$

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

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

choose α_i if $R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$

Losses and Risks: 0/1 Loss

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

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

For minimum risk, choose the most probable class

Losses and Risks: Reject

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

$$R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$$

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

choose C_i if $P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x}) \; \forall k \neq i$ and $P(C_i \mid \mathbf{x}) > 1 - \lambda$

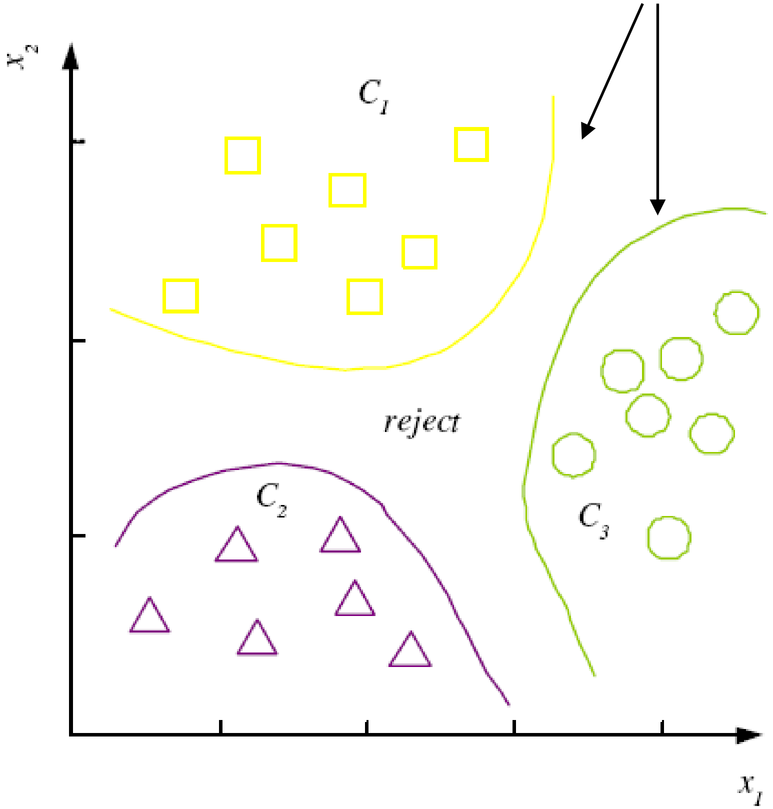
reject otherwise

Discriminant Functions

choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

$g_i(\mathbf{x}), i = 1, \ldots, K$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i \mid \mathbf{x}) \\ P(C_i \mid \mathbf{x}) \\ p(\mathbf{x} \mid C_i) P(C_i) \end{cases}$$



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$