ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`I at Manoa

Kyungim Baek

Announcement

- Homework assignment # 1 has been posted
  - Due: **Monday September 21, by 5:00 PM**
- Project proposal
  - Due: **Wednesday September 23, by 5:00 PM**

Lecture 3

- Bayesian Decision Theory
  - Minimum Risk Classification
  - Minimum Error-Rate Classification
- Discriminant Functions

Previously…

- Single feature, two categories
- **Bayes decision rule**

  Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2
  - Equivalently:

  Decide ω_1 if $p(x|\omega_1)\cdot P(\omega_1) > p(x|\omega_2)\cdot P(\omega_2)$; otherwise decide ω_2
- It minimize the average probability of error

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$

where

$P(error|x) = P(\omega_1|x)$ if we decide ω_2

$P(error|x) = P(\omega_2|x)$ if we decide ω_1

Bayesian Decision Theory: Generalization

- Generalization of the preceding ideas by
  - Use of more than one feature
  - Use more than two categories
  - Allowing actions and not only decide on the category
    - Allows the possibility of rejection; i.e. refusing to make a decision in close or bad cases!
  - Introducing a loss of function which is more general than the probability of error
    - The ***loss function*** states how costly each action taken is

Generalization (cont'd)

- Let
  - $\{\omega_1, \omega_2, \ldots, \omega_c\}$: set of c classes
  - $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$: set of a possible actions
  - $\lambda(\alpha_i | \omega_j)$: loss incurred for taking action α_i when the true class is ω_j
- ***Conditional risk***: expected loss (i.e. risk) associated with taking action α_i, given the observation **x**

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

Generalization (cont'd)

- Overall risk: expected loss associated with a given *decision rule* $\alpha(\mathbf{x})$, specifying the action

$$R = \int R(\alpha(\mathbf{x}) | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- Minimizing R
  $\Leftrightarrow$ Selecting α_i with minimum conditional risk $R(\alpha_i | \mathbf{x})$ (resulting R is called the ***Bayes risk***)
- ***Bayes decision rule***:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x)$$

$$\alpha(\mathbf{x}) = \arg\min_{1 \le i \le a} R(\alpha_i|\mathbf{x})$$

Two-Category Classification

- α_1: deciding ω_1; α_2: deciding ω_2
  $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$: loss incurred for deciding ω_i when the true class is ω_j
- Conditional risk: $\boxed{R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x)}$

$$R(\alpha_1 | \mathbf{x}) = \lambda_{11}P(\omega_1 | \mathbf{x}) + \lambda_{12}P(\omega_2 | \mathbf{x})$$
$$R(\alpha_2 | \mathbf{x}) = \lambda_{21}P(\omega_1 | \mathbf{x}) + \lambda_{22}P(\omega_2 | \mathbf{x})$$

- Our rule is the following:
  if $R(\alpha_1 | \mathbf{x}) < R(\alpha_2 | \mathbf{x})$, take action α_1 (decide ω_1)
- This results in the equivalent rule:

$$\text{decide} \begin{cases} \omega_1 & \text{if } (\lambda_{21} - \lambda_{11})p(\mathbf{x} | \omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x} | \omega_2)P(\omega_2) \\ \omega_2 & \text{otherwise} \end{cases}$$

Two-Category Classification

- **Likelihood ratio:**
  – The preceding rule is equivalent to the following rule:

$$\begin{cases} \text{take action } \alpha_1 \text{ (decide } \omega_1) & \text{if } \dfrac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \dfrac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \dfrac{P(\omega_2)}{P(\omega_1)} \\ \text{take action } \alpha_2 \text{ (decide } \omega_2) & \text{otherwise} \end{cases}$$

- Optimal decision property
  – "If the likelihood ratio exceeds a threshold value that is independent of the input pattern **x**, we can take optimal actions"

Likelihood Ratio Decision Rule: Illustration of Simple Case

- Assume that $\lambda_{ii} = 0$ and $P(\omega_1) = P(\omega_2)$. Then, the decision rule becomes

$$\begin{cases} \text{take action } \alpha_1 \text{ (decide } \omega_1) & \text{if } \dfrac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \dfrac{\lambda_{12}}{\lambda_{21}} \\ \text{take action } \alpha_2 \text{ (decide } \omega_2) & \text{otherwise} \end{cases}$$

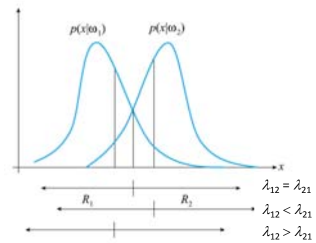


Figure from *PR* by Ilsuk Oh

Minimum-Error-Rate Classification

- Actions are decisions on classes:
  – If action α_i is taken and the true class is ω_j then the decision is correct if $i = j$ and in error if $i \neq j$
- Seek a decision rule that minimizes the ***probability of error*** which is the ***error rate***
- The *symmetrical* or *zero-one loss* function:

$$\lambda(\alpha_i \,|\, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

  – No loss to a correct decision
  – Unit loss to any error, i.e. the errors are equally costly

Minimum-Error-Rate Classification

- Conditional risk corresponding to zero-one loss function is the average probability of error:

$$R(\alpha_i \,|\, \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \,|\, \omega_j) P(\omega_j \,|\, \mathbf{x})$$
$$= \sum_{j \neq i} P(\omega_j \,|\, \mathbf{x})$$
$$= 1 - P(\omega_i \,|\, \mathbf{x})$$

- Minimize the risk requires maximize $P(\omega_i|\mathbf{x})$; i.e. for minimum error rate, **decide ω_i if $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \;\; \forall j \neq i$**
  – In other words, select class maximizing the posterior probability (as recommended by the Bayes decision rule)!

Minimum-Error-Rate Classification

- Regions of decision and zero-one loss function

Let $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda$ then decide ω_1 if $\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \theta_\lambda$

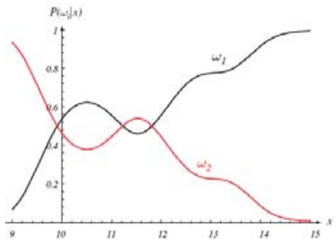
- If λ is the zero-one loss function which means:

$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

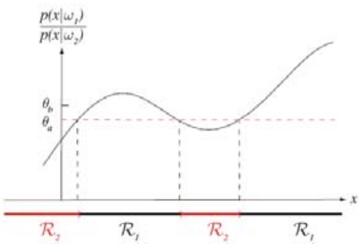
$$\text{then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

Minimum-Error-Rate Classification

- Regions of decision and zero-one loss function



Posterior probabilities for particular priors $P(\omega_1)$ = 2/3 and $P(\omega_2)$ = 1/3



Likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the case on the left.

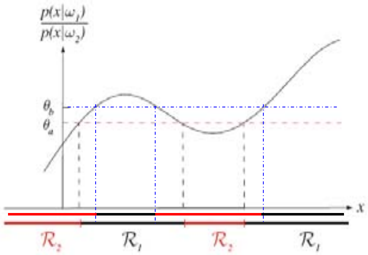
Unequal Loss

- For example, $\lambda_{12} = 2 \times \lambda_{21}$, i.e.

$$\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$$

then

$$\theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

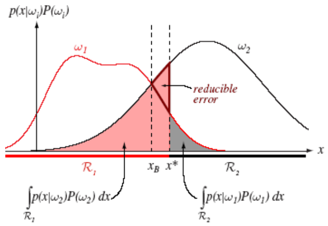


If misclassifying ω_2 as ω_1 is penalized more than the converse, the threshold gets larger (θ_b), and hence R_1 becomes smaller.

Suboptimal Decisions

$$\begin{aligned} P(error) &= \int P(error, \mathbf{x}) d\mathbf{x} \\ &= P(\mathbf{x} \in R_2, \omega_1) + P(\mathbf{x} \in R_1, \omega_2) \\ &= P(\mathbf{x} \in R_2 \mid \omega_1) P(\omega_1) + P(\mathbf{x} \in R_1 \mid \omega_2) P(\omega_2) \\ &= \int_{R_2} p(\mathbf{x} \mid \omega_1) P(\omega_1) d\mathbf{x} + \int_{R_1} p(\mathbf{x} \mid \omega_2) P(\omega_2) d\mathbf{x} \end{aligned}$$

If the decision point (x^*) is selected arbitrarily, then the error is not as small as it should be. The area of "reducible error" can be eliminated by moving the decision point to the Bayes optimal decision boundary (x_B). In conclusion, Bayes decision rule maximizes the probability of correct classification and minimizes the risk of error.



Summary

- **Bayes decision rule** $\alpha(\mathbf{x})$
  - Conditional risk:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

  - Overall risk:

$$R = \int R(\alpha(\mathbf{x})\,|\,\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

  - Minimize the overall risk: $\alpha(\mathbf{x}) = \arg\min_{1\le i\le a} R(\alpha_i|\mathbf{x})$

Summary (cont'd)

- **Bayes decision rule** $\alpha(\mathbf{x})$
  - If λ is the *zero-one loss* function: $\lambda(\alpha_i\,|\,\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \ne j \end{cases} \quad i, j = 1, \ldots, c$

$$R(\alpha_i\,|\,\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i\,|\,\omega_j)P(\omega_j\,|\,\mathbf{x})$$
$$= \sum_{j\ne i} P(\omega_j\,|\,\mathbf{x})$$
$$= 1 - P(\omega_i\,|\,\mathbf{x})$$

Then, $\alpha(\mathbf{x}) = \arg\min_{1\le i\le a} R(\alpha_i|\mathbf{x}) = \omega_i$ if $P(\omega_i\,|\,\mathbf{x}) > P(\omega_j\,|\,\mathbf{x}) \quad \forall j \ne i$

- In other words, select the class maximizing the posterior probability
- Minimize the probability of error (i.e. minimum error rate)

Classifiers, Discriminant Functions and Decision Surfaces

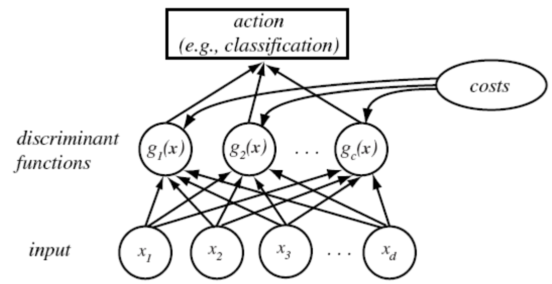
- **Discriminant function**: a function employed for differentiating/discriminating between classes
  - Pattern classifiers can be represented by set of discriminant functions, $g_i(\mathbf{x})$, $i = 1, \ldots, c$
  - Decision rule:

Assign a feature vector $\mathbf{x}$ to class i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \ne i$$

Generic Classifier

- Functional structure of a general statistical pattern classifier



Discriminant Functions for Bayes Classifier

- General case (minimum conditional risk):

  $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$

- Minimum error rate (maximum posterior):

  $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$

  or equivalently,

  $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$

  $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$ (ln: natural logarithm)

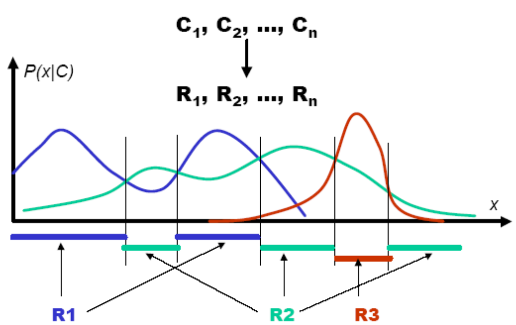
Decision Regions

- A decision rule based on the discriminant functions generates a set of ***decision regions*** (i.e. divide feature space into c decision regions), $R_1, R_2, \ldots, R_c$ (not needing to be contiguous):

  if $g_i(\mathbf{x}) > g_j(\mathbf{x})\ \ \forall j \neq i$ then $\mathbf{x}$ is in R_i

- Decision regions are separated by ***decision boundaries***. The condition satisfied at the decision boundary between R_i and R_j is: $g_i(\mathbf{x}) = g_j(\mathbf{x})$

Decision Regions and Boundaries



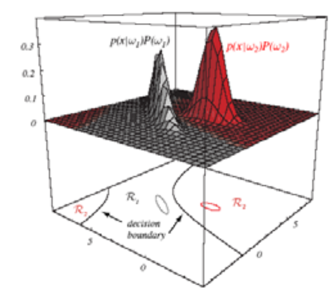
from S. Iliescu

The Two-Category Case

- A classifier is a “dichotomizer” that has two discriminant functions g_1 and g_2

  Let $g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$

  $$\text{Decide} \begin{cases} \omega_1 & \text{if } g(\mathbf{x}) > 0 \\ \omega_2 & \text{otherwise} \end{cases}$$



- Example: $g(\mathbf{x})$ for minimum error-rate classification

  $$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

  $$g(\mathbf{x}) = \ln\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}$$