Chapter 2 (Part 1): Bayesian Decision Theory (Sections 2.1-2.2)

Bayesian Decision Theory – Continuous Features…

- Let $\{\omega 1, \omega 2, \ldots, \omega c\}$ be the set of c states of nature (or "categories")
- Let $\{\alpha 1, \alpha 2, \ldots, \alpha a\}$ be the set of possible actions
- Let $\lambda(\alpha i \mid \omega j)$ be the loss incurred for taking action αi when the true state of nature is ωj

Bayesian Decision Theory – Continuous Features…

- Overall risk

- R = Sum of all $\underbrace{R(\alpha i \mid x)}_{\text{Conditional risk}}$ for i = 1,…,a

- Minimizing R ⟺ Minimizing $R(\alpha i \mid x)$ for i = 1,…, a

- $$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$ for i = 1,…,a

Bayesian Decision Theory – Continuous Features…

- Select the action αi for which $R(\alpha i \mid x)$ is minimum

$$\Downarrow$$

- R is minimum and is called the Bayes risk = best performance that can be achieved!

Two-category Classification

- α_1 : deciding ω_1
- α_2 : deciding ω_2
- $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Loss incurred for deciding α_i when the true state of nature is ω_j
- Conditional risk:
  - $R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$
  - $R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$

Two-category Classification

- Our rule is the following:

  if $R(\alpha 1 \mid x) < R(\alpha 2 \mid x)$

- Action $\alpha 1$: "decide $\omega 1$" is taken
- This results in the equivalent rule :
- Decide $\omega 1$ if: $(\lambda 21 - \lambda 11)\, P(x \mid \omega 1)\, P(\omega 1) >$

  $(\lambda 12 - \lambda 22)\, P(x \mid \omega 2)\, P(\omega 2)$

- and decide $\omega 2$ otherwise

Likelihood Ratio

- The preceding rule is equivalent to the following rule:

$$if\ \frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

- Then take action $\alpha 1$ (decide $\omega 1$)
- Otherwise take action $\alpha 2$ (decide $\omega 2$)

Likelihood Ratio…

- Optimal decision property

"If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions"

Example: Checking on a course

- A student needs to make a decision which courses to take, based only on first lecture's impression
- From student's previous experience:

Quality of the course	good	fair	bad
Probability (prior)	0.2	0.4	0.4

- These are **prior probabilities**.

Example is taken from webcourse.cs.technion.ac.il/236875/Spring2007/ho/WCFiles/Tutorial2.ppt

Example: Checking on a course

- **The student also knows the class-conditionals:**

$Pr(x/w_j)$	good	fair	bad
Interesting lecture	0.8	0.5	0.1
Boring lecture	0.2	0.5	0.9

- **The loss function is given by the matrix**

$l(a_i/w_j)$	good course	fair course	bad course
Taking the course	0	5	10
Not taking the course	20	5	0

Example: Checking on a course

- **The student also knows the class-conditionals:**

$Pr(x/w_j)$	good	fair	bad
Interesting lecture	0.8	0.5	0.1
Boring lecture	0.2	0.5	0.9

- **The loss function is given by the matrix**

$\lambda(\alpha i \mid \omega j)$	good course	fair course	bad course
Taking the course	0	5	10
Not taking the course	20	5	0

Example: Checking on a course

- We can take decisions according to two rules:
  - Posterior probability
  - Risk minimization

Example: Checking on a course

- Given that the first class was interesting, the student wants to make an optimal decision.
  - Pr(bad|interesting)
  - Pr(fair|interesting)
  - Pr(good|interesting)

  - Pr(interesting|bad)
  - Pr(interesting|fair)
  - Pr(interesting|good)

  - Pr(interesting)

Example: Checking on a course

- The probability to get the "interesting lecture"(x= interesting):
- Pr(interesting)= Pr(interesting|good course)* Pr(good course)

  + Pr(interesting|fair course)* Pr(fair course)

  + Pr(interesting|bad course)* Pr(bad course)

  =0.8*0.2+0.5*0.4+0.1*0.4=0.4
- Consequently, Pr(boring)=1-0.4=0.6
- Suppose the lecture was interesting. Then we want to compute the **posterior** probabilities of each one of the 3 possible "states of nature".

Example: Checking on a course

$$\Pr(\text{good course}|\text{interesting lecture}) = \frac{\Pr(\text{interesting}|\text{good})\Pr(\text{good})}{\Pr(\text{interesting})} = \frac{0.8*0.2}{0.4} = 0.4$$

$$\Pr(\text{fair}|\text{interesting}) = \frac{\Pr(\text{interesting}|\text{fair})\Pr(\text{fair})}{\Pr(\text{interesting})} = \frac{0.5*0.4}{0.4} = 0.5$$

- We can get Pr(bad|interesting) = 0.1 either by the same method, or by noting that it complements the above two to 1.
- Now, we have all we need to make an intelligent decision about an optimal action

Example: Checking on a course

- The student needs to minimize the conditional risk

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

R(taking|interesting) =

Pr(good|interesting) λ(taking good course) + Pr(fair|interesting) λ(taking fair course) + Pr(bad|interesting) λ(taking bad course)

= 0.4*0+0.5*5+0.1*10 = 3.5

Example: Checking on a course

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

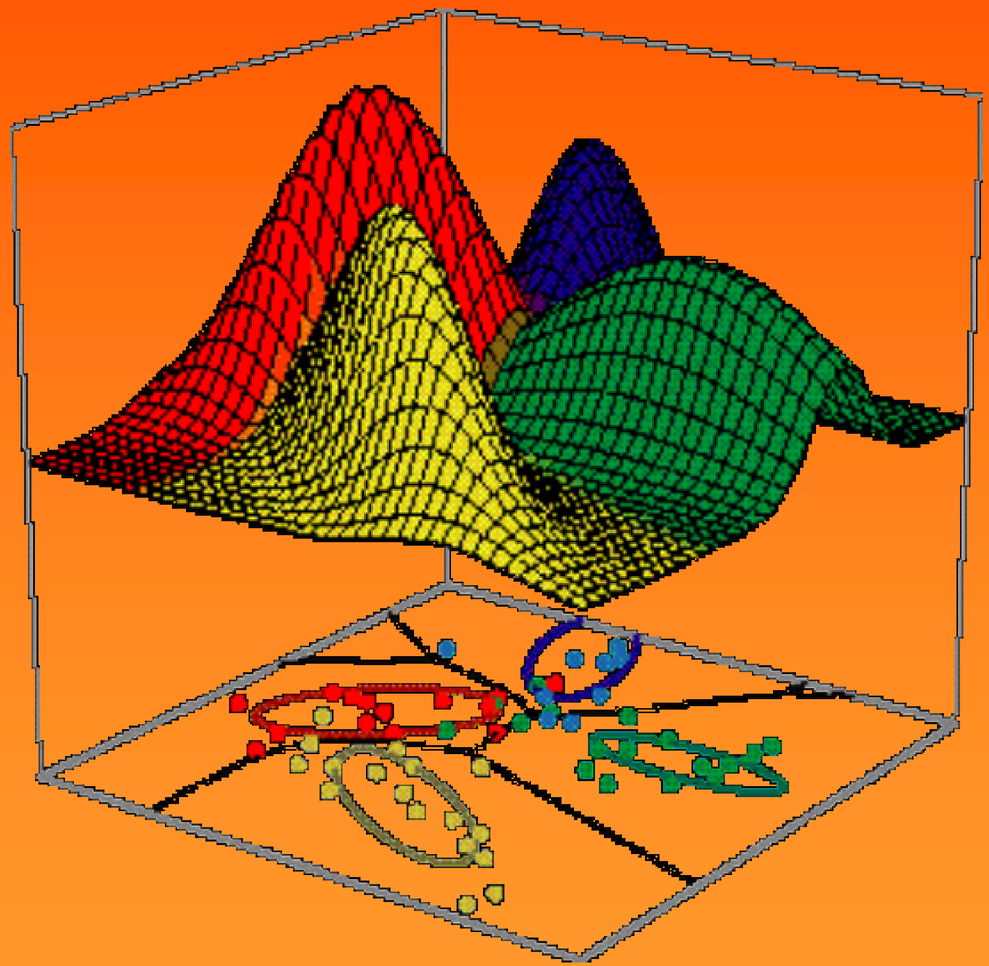
R(not taking|interesting) =

Pr(good|interesting)λ(not taking good course) +
Pr(fair|interesting)λ(not taking fair course) +
Pr(bad|interesting)λ(not taking bad course)

= 0.4*20+0.5*5+0.1*0 = 10.5

Constructing an optimal decision function

- So, if the first lecture was interesting, the student will minimize the conditional risk <u>by taking the course</u>.
- In order to construct the full decision function, we need to define the risk minimization action for the case of boring lecture, as well.



Pattern Classification

All materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000 with the permission of the authors and the publisher