Decision Theory

Chris Williams

School of Informatics, University of Edinburgh

October 2010

Overview

- Classification and Bayes decision rule
- Sampling vs diagnostic paradigm
- Classification with Gaussians
- Loss, Utility and Risk
- Reject option
- Reading: Bishop §1.5

Classification

How should we assign example $\mathbf{x}$ to a class $\mathcal{C}_k$?

1. use discriminant functions $y_k(\mathbf{x})$
2. model class-conditional densities $P(\mathbf{x}|\mathcal{C}_k)$ and then use Bayes' rule
3. Model posterior probabilities $P(\mathcal{C}_k|\mathbf{x})$ directly

Approaches 2 and 3 give a two-step decision process

- *Inference* of $P(\mathcal{C}_k|\mathbf{x})$
- *Decision making* in the face of uncertainty

- Bayes decision rule: allocate example $\mathbf{x}$ to class k if

$$P(\mathcal{C}_k|\mathbf{x}) > P(\mathcal{C}_j|\mathbf{x}) \qquad \forall j \neq k$$

- This rule minimizes the expected error at $\mathbf{x}$. Proof: Choosing class i will lead to

$$P(\text{error}|\mathbf{x}) = 1 - P(\mathcal{C}_i|\mathbf{x})$$

  This is minimized by choosing $i = k$. Note that a randomized allocation rule is not superior.
- Using Bayes' rule, rewrite decision rule as

$$P(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) > P(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j) \qquad \forall j \neq k$$

- $P(\text{error})$ is minimized by this decision rule

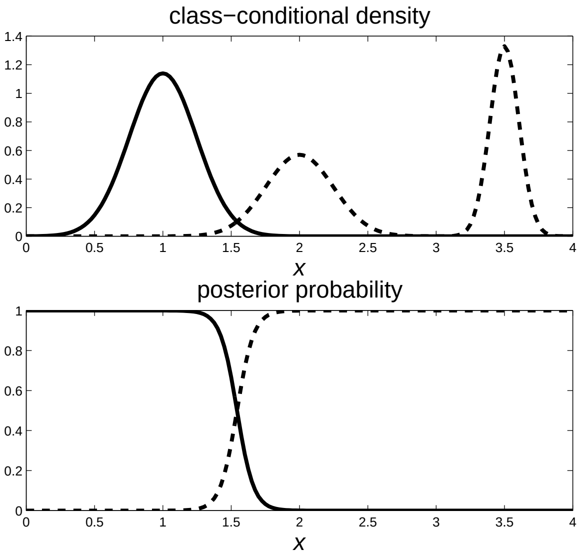
$$\begin{aligned} P(\text{error}) &= \int P(\text{error}, \mathbf{x})\, d\mathbf{x} \\ &= \int P(\text{error}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

Errors in classification arise from

1. Errors due to class overlap
   these are unavoidable
2. Errors resulting from an incorrect decision rule
   use the correct rule!
3. Errors resulting from an inaccurate model of the posterior probabilities
   accurate modelling is a challenging problem

Model $P(\mathcal{C}_k|\mathbf{x})$ or $P(\mathbf{x}|\mathcal{C}_k)$?

- Diagnostic paradigm (discriminative): Model $P(\mathcal{C}_k|\mathbf{x})$ directly
- Sampling paradigm (generative): Model $P(\mathbf{x}|\mathcal{C}_k)$ and $P(\mathcal{C}_k)$
- Pros/cons of diagnostic paradigm:
  - ☺ Modelling $P(\mathcal{C}_k|\mathbf{x})$ can be simpler than modelling class-conditional densities.
  - ☺ Less sensitive to modelling assumptions as what we need, $P(\mathcal{C}_k|\mathbf{x})$ is modelled directly
  - ☹ Marginal density $p(\mathbf{x})$ is needed to handle outliers and missing values
  - ☹ Use of unclassified observations difficult in diagnostic paradigm
  - ☹ Dealing with missing inputs is difficult


class−conditional density
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x
posterior probability
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x

Classification with Gaussians

- Check if
$$\frac{P(\mathcal{C}_1|\mathbf{x})}{P(\mathcal{C}_2|\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \gtrless 1$$
or if
$$\Delta(\mathbf{x}) = \log \frac{p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \gtrless 0$$
- For Gaussian class-conditional densities and $\Sigma_1 = \Sigma_2$ we obtain
$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1) + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \gtrless 0$$
This is a *linear* classifier
- For $\Sigma_1 \neq \Sigma_2$, boundaries are hyperquadrics

Loss and Risk

- Actions $a_1, \ldots, a_A$ might be taken. Given $\mathbf{x}$, which one should be taken?
- L_{ji} is the loss incurred if action a_i is taken when the state of nature is $\mathcal{C}_j$
- The expected loss (or risk) of taking action a_i given $\mathbf{x}$ is

$$R(a_i|\mathbf{x}) = \sum_j L_{ji} P(\mathcal{C}_j|\mathbf{x})$$

- Choose action k if

$$\sum_j L_{jk} P(\mathcal{C}_j|\mathbf{x}) < \sum_j L_{ji} P(\mathcal{C}_j|\mathbf{x}) \qquad \forall i \neq k$$

- Let $a(\mathbf{x}) = \mathrm{argmin}_i R(a_i|\mathbf{x})$
- Overall risk R

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

Example loss function

- Patients are classified to classes $\mathcal{C}_1 =$ healthy, $\mathcal{C}_2 =$ tumour.
- Actions are $a_1 =$ discharge the patient, $a_2 =$ operate
- Assume $L_{11} = L_{22} = 0$, $L_{12} = 1$ and $L_{21} = 10$, i.e. it is 10 times worse to discharge the patient when they have a tumour than to operate when they do not

$$R(a_1|\mathbf{x}) = L_{11}P(\mathcal{C}_1|\mathbf{x}) + L_{21}P(\mathcal{C}_2|\mathbf{x}) = L_{21}P(\mathcal{C}_2|\mathbf{x})$$
$$R(a_2|\mathbf{x}) = L_{12}P(\mathcal{C}_1|\mathbf{x}) + L_{22}P(\mathcal{C}_2|\mathbf{x}) = L_{12}P(\mathcal{C}_1|\mathbf{x})$$

- Choose action a_1 when $R(a_1|\mathbf{x}) < R(a_2|\mathbf{x})$, i.e. when

$$L_{21}P(\mathcal{C}_2|\mathbf{x}) < L_{12}P(\mathcal{C}_1|\mathbf{x})$$

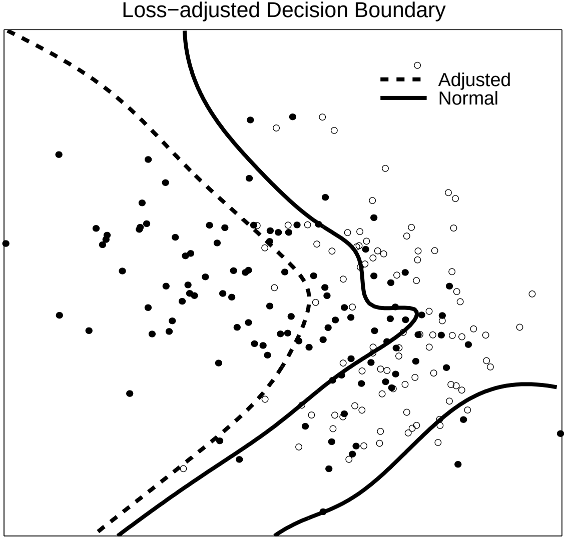
or

$$\frac{P(\mathcal{C}_1|\mathbf{x})}{P(\mathcal{C}_2|\mathbf{x})} > \frac{L_{21}}{L_{12}} = 10$$

- If $L_{21} = L_{12} = 1$ then threshold is 1; in our case we require stronger evidence in favour of $\mathcal{C}_1 =$ healthy in order to discharge the patient

- In credit risk assignment, losses are monetary
- Note that rescaling loss matrix does not change the decision
- Minimum classification error is obtained by

$$L_{ji} = 1 - \delta_{ji}$$


Loss−adjusted Decision Boundary
Adjusted
Normal

Utility and Loss

- Basically same thing with opposite sign. Maximize expected utility, minimize expected loss.
- See Russell and Norvig ch 16 for a discussion of fundamentals of utility theory, and utility of money [not examinable]
- Russell and Norvig ch 17 discuss sequential decision problems. Involves utilities, uncertainty and sensing; generalizes problems of planning and search. See RL course.

Reject option

$$P(\mathrm{error}|\mathbf{x}) = 1 - \max_j P(\mathcal{C}_j|\mathbf{x})$$

- If we can reject some examples, reject those that are most confusable, i.e. where $P(\mathrm{error}|\mathbf{x})$ is highest
- Choose a threshold θ and reject if

$$\max_j P(\mathcal{C}_j|\mathbf{x}) < \theta$$

- Gives rise to error-reject curves as θ is varied from 0 to 1

Error-reject curve