Bayes decision theory

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Machine Learning
CS230
Sem 3, 2018-19

Introduction

- A probabilistic approach to classification problems.
- The theory enables optimal decision making in a probabilistic setting.
- Idea: Select the class for which the expected risk is the least.
  - Generally, the risk incorporates the costs linked with different decisions.
- Problem needs to formulated in a probabilistic framework, and all relevant probabilities are assumed to be known.

Bayes rule in classification problems

- Enables computation of the posterior probability as

$$\boxed{P(y|\mathbf{x}) = \frac{p(\mathbf{x}|y) \times P(y)}{p(\mathbf{x})}}$$

  - $P(y|\mathbf{x})$: Probability of the output given a particular input.
  - $p(\mathbf{x}|y)$: Probability of the input data given a particular output.
  - $P(y)$: Prior probability of the output (class), without observing the data.
  - $p(\mathbf{x})$: Probability of the input observation.

Example

- Classification of public transport.

Auto

Class c_1

Taxi

Class c_2

Bus

Class c_3

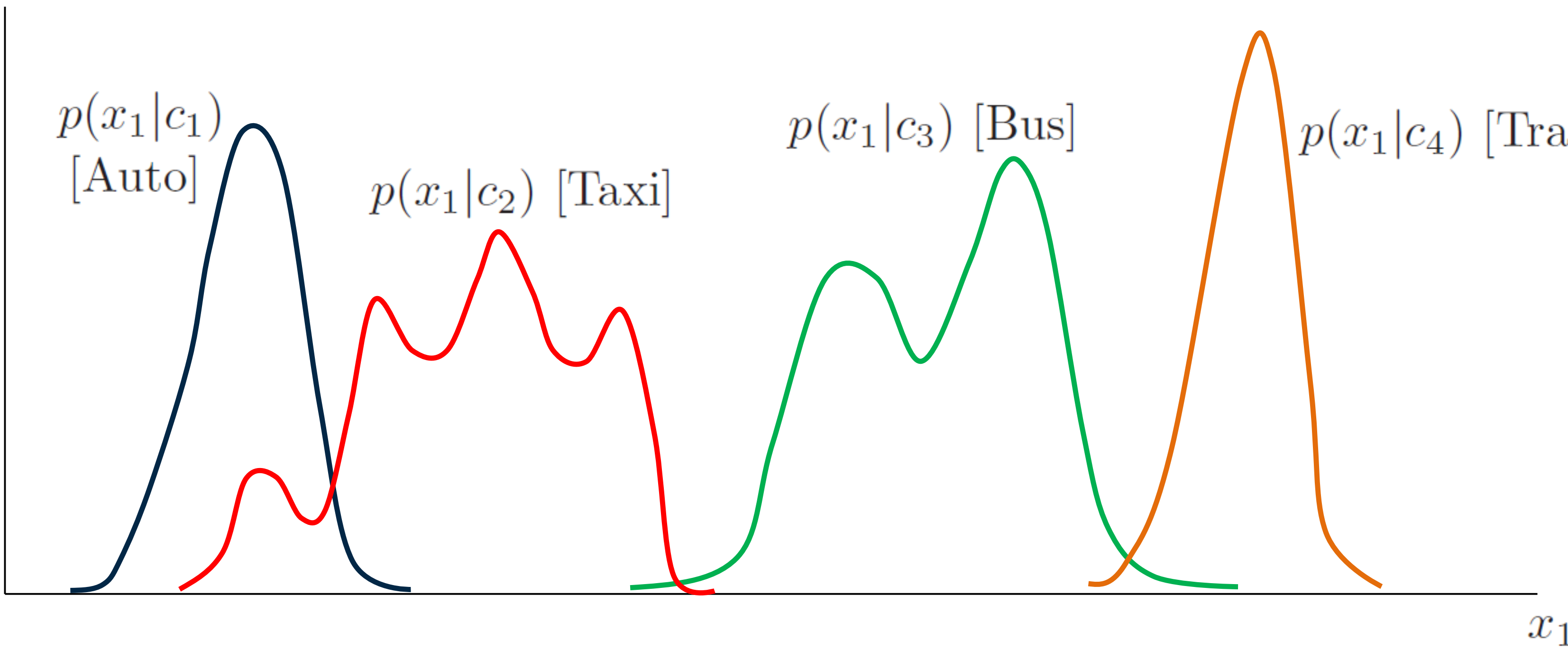
Tram

Class c_4

- Features:
  - Length (x_1)
  - Width (x_2)
  - Height (x_3)
  - Weight (x_4)

Class-conditional probability density

- It is the probability density function for feature $\mathbf{x}$ given a particular class, e.g. $p(\mathbf{x}|c_2)$.
- This is the class likelihood.
- Example: Hypothetical class-conditional probability density for the first feature, length (x_1), of the four classes.



Prior

- Prior probability reflects the a priori knowledge of the outputs (classes) before the feature observations are taken into account.

Training Dataset

- Prior probabilities

$\boxed{\text{Auto: } P(c_1) = \frac{5}{23}}$ $\boxed{\text{Taxi: } P(c_2) = \frac{4}{23}}$ $\boxed{\text{Bus: } P(c_3) = \frac{8}{23}}$ $\boxed{\text{Tram: } P(c_4) = \frac{6}{23}}$

- For J classes the priors must satisfy

$$\sum_{j=1}^{J} P(c_j) = 1$$

Posterior

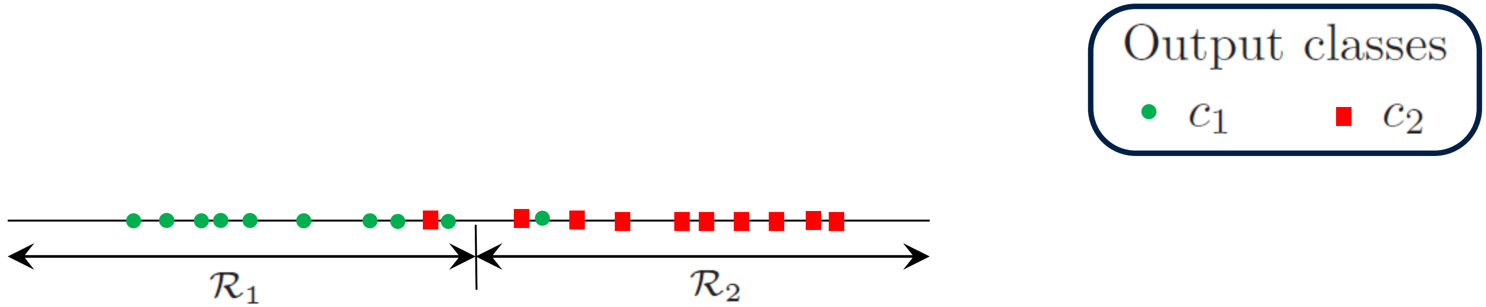
- Posterior probability is the probability of an output (say the jth class) given some input $\mathbf{x}$:

$$P(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)P(c_j)}{p(\mathbf{x})}$$

- The term $p(\mathbf{x})$ is constant for all classes and as such can be ignored.

- Thus, the class-conditional probability density $p(\mathbf{x}|c_j)$ and the prior $P(c_j)$ govern the posterior probability $P(c_j|\mathbf{x})$.

Decision boundary



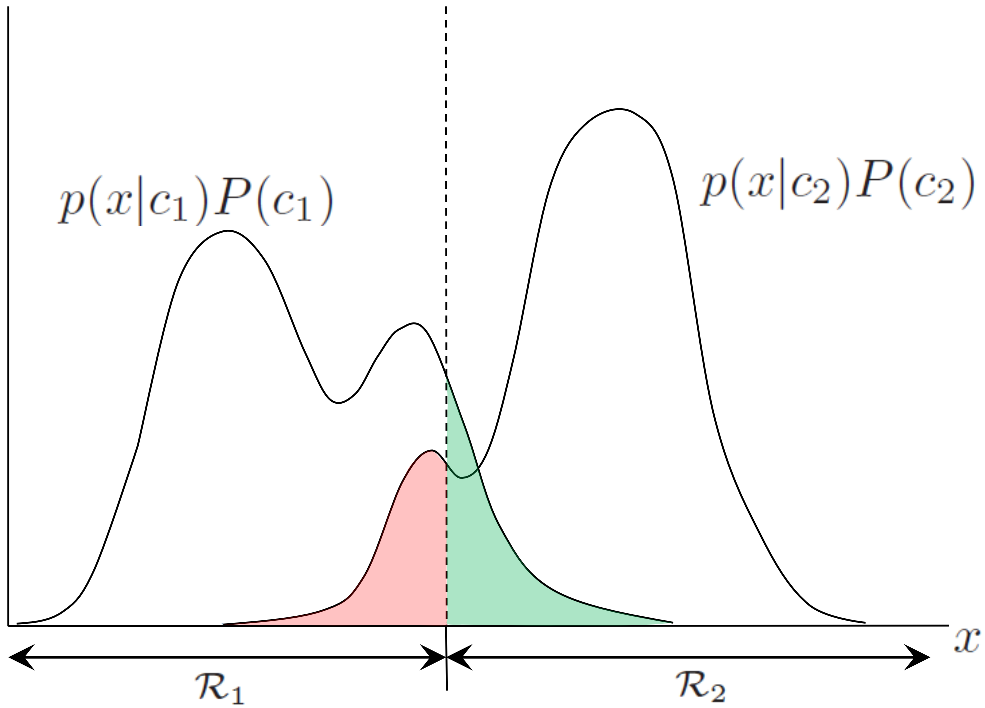
- Consider a simple dataset:
  - 1D feature space
  - Two classes

- In this case there are two ways in which data can be misclassified:
  - x belongs to c_1, but is located in decision region $\mathcal{R}_2$.
  - x belongs to c_2, but is located in decision region $\mathcal{R}_1$.

- The probability of error given x:

$$P(\text{error}|x) = \begin{cases} P(c_2|x) & \text{if } x \text{ is assigned to } c_1 \\ P(c_1|x) & \text{if } x \text{ is assigned to } c_2 \end{cases}$$

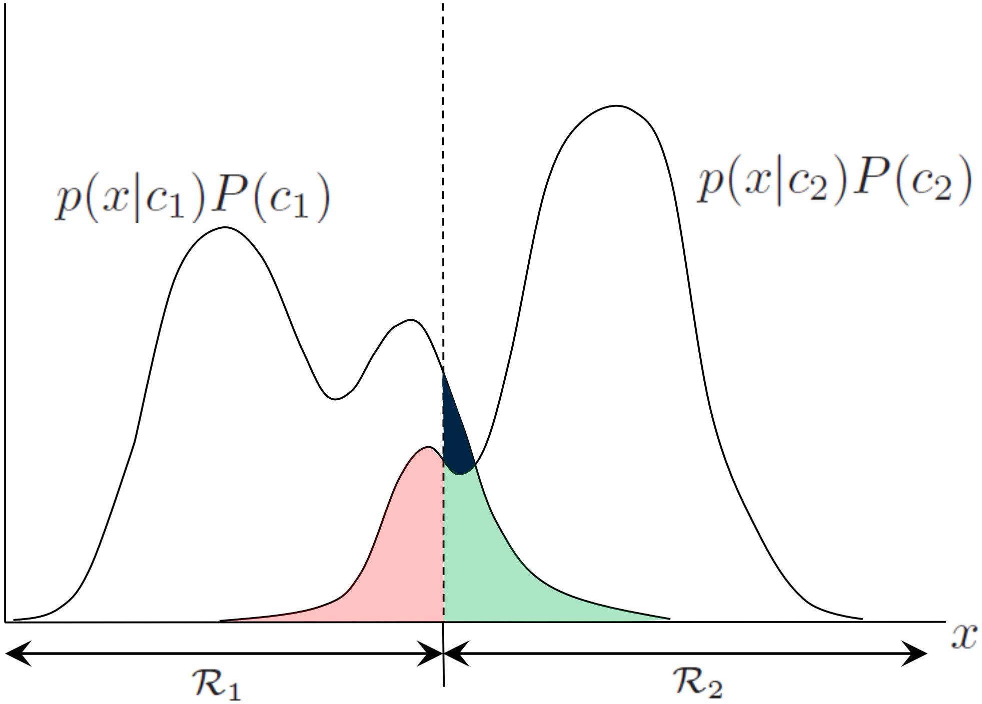
Bayes Error



- Average probability of error:

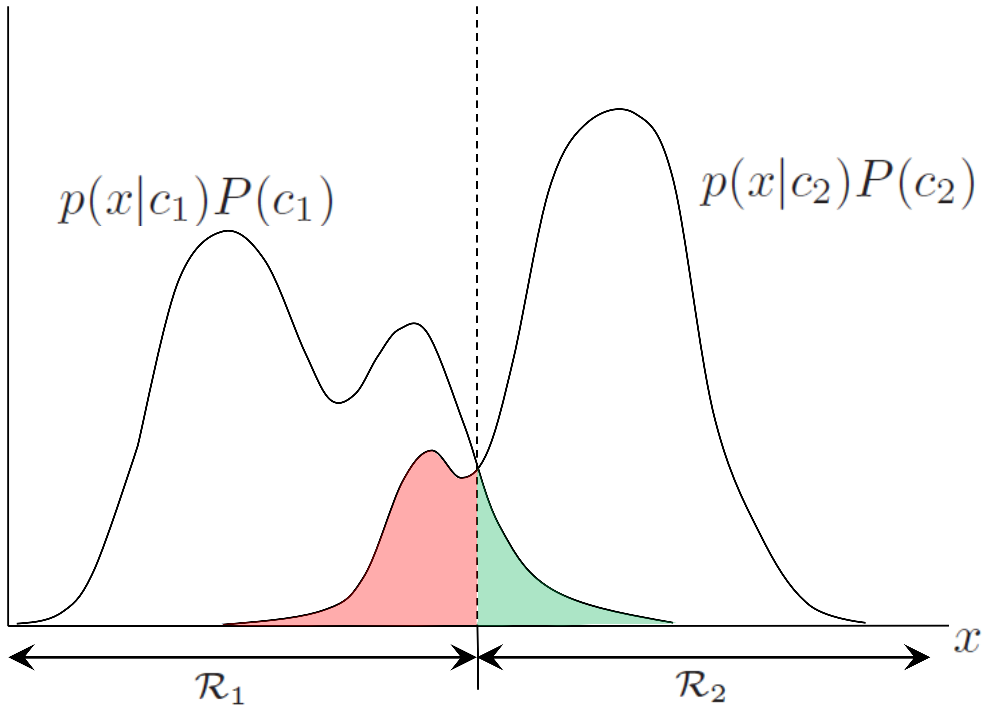
$$\begin{aligned} P(\text{error}) &= \int_{-\infty}^{\infty} P(\text{error}|x)p(x)\mathrm{d}x \\ &= \int_{\mathcal{R}_1} P(c_2|x)p(x)\mathrm{d}x + \int_{\mathcal{R}_2} P(c_1|x)p(x)\mathrm{d}x \\ &= \int_{\mathcal{R}_1} p(x|c_2)P(c_2)\mathrm{d}x + \int_{\mathcal{R}_2} p(x|c_1)P(c_1)\mathrm{d}x \end{aligned}$$

Reducible Error



- Reducible Error: Error produced due to suboptimal choice of decision boundary.

Bayes decision rule



- Probability of misclassification is the least when each data point is assigned to the class with maximum posterior probability $P(c_j|x)$.

General theory

- A risk function (more general form of error function) is derived from the losses incurred from all the errors.
- Suppose there are J output classes – $\{c_1, c_2, ..., c_J\}$.
- The loss function computes the cost of taking an action.
- Let $L(\alpha_i|c_j)$ the cost of taking action α_i when the actual class is c_j.
- In the simplest case, actions could be same as the classes, i.e. $\alpha_i = c_i$.
- Let $R(\alpha_i|\mathbf{x})$ be the expected loss or conditional risk of taking action α_i for a particular input $\mathbf{x}$, and is defined as

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{J} L(\alpha_i|c_j)P(c_j|\mathbf{x})$$

General theory

- $R(\alpha_i|\mathbf{x})$ is expected (average) loss for taking an action for a particular input and loss function.
- If actions and classes are the same, then $\alpha_i = c_i$.
- The overall risk of a decision rule is the expected loss associated with a given decision rule:

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

- In order to minimize the overall risk, we need a rule that minimizes $R(\alpha_i|\mathbf{x})$ for all $\mathbf{x}$.
- The Bayes decision rule minimizes the overall risk by selecting the action that minimizes the conditional risk:

$$\begin{aligned}\alpha^* &= \arg\min_{\alpha_i} R(\alpha_i|\mathbf{x}) \\ &= \arg\min_{\alpha_i} \sum_{j=1}^{J} L(\alpha_i|c_j)P(c_j|\mathbf{x})\end{aligned}$$

Zero-one loss function

- The Zero-One Loss function is widely used and is defined as

$$L(\alpha_i|c_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

  where $i, j = 1, 2, ..., J$.
  - There is no loss for taking correct decision.
  - Incorrect decisions incur uniform unit loss.

- The conditional risk in this case becomes

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{J} L(\alpha_i|c_j) P(c_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(c_j|\mathbf{x}) \\ &= 1 - P(c_i|\mathbf{x}) \end{aligned}$$

- Therefore, for a particular $\mathbf{x}$, the conditional risk is minimized by taking the action α_i that maximizes $P(c_i|\mathbf{x})$.