Foundations of Machine Learning
AI2000 and AI5000

FoML-11
Bayesian Regression

FoML-12
Decision Theory

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad

July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

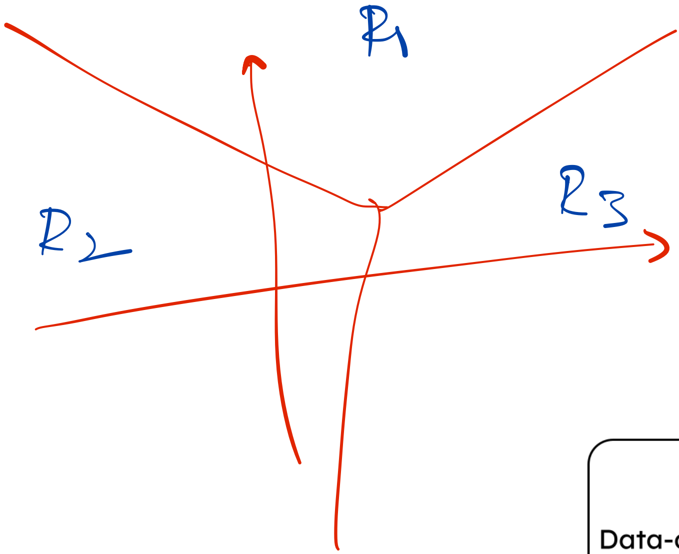
So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Linear Regression with basis functions - and regularization
- Model selection
- Bias-Variance Decomposition/Tradeoff (Bayesian Regression)

Decision Theory

Decision Theory

- Dataset: i/p vectors $\mathbf{x} \in \mathbb{R}^D$, ground truth $t \in \{C_1, C_2, \ldots, C_K\}$
- Divide the i/p space $\mathbb{R}^D$ into K decision regions $R_k,\ k = \{1, 2, \ldots K\}$
- For every data point
  - Ground truth t_n
  - Prediction $y(\underline{x}_n, \underline{w}) = \hat{t}_n$



Decision Theory

- Confusion Matrix

$$
\begin{array}{c c c}
 & & R_1 \quad R_2 \quad \cdots \quad R_K \quad \text{prediction (By the model)} \\
 & c_1 & \\
\text{GT} & c_2 & \begin{bmatrix} 6 & 2 & 1 & 0 & 0 & 0 \\ 1 & 5 & 0 & 0 & 2 & 1 \\ \vdots & & & & & \vdots \\ 1 & 0 & 0 & 0 & 0 & 7 \end{bmatrix} \\
 & \vdots & \\
 & c_K &
\end{array}
$$

Diagonal elements - correct predictions

Off-diagonal elements - incorrect predictions

Decision Theory - Misclassification Rate

- Goal of classification - Minimize the misclassification rate
- Assume the data are drawn independently from the joint distribution
- Probability of a misclassification: $p(\text{mistake}) = \sum_{i=1}^{K} \sum_{k \neq i} p(\mathbf{x} \in R_i, C_k)$

$p(\text{mistake}) =$ 1 - P(Correct classification)

$= 1 - \sum_{k=1}^{K} P(x \in R_k, C_k)$

Decision Theory - Misclassification Rate

- Minimizing the misclassification rate (how to ensure this?)
  - Assign x to class C_k if $p(\mathbf{x}, t = C_k) > p(\mathbf{x}, t = C_j), \ \forall j \neq k$
  - We know that

$$p(\underline{x}, C_k) = \underbrace{p(C_k|\underline{x})}_{\text{look for the largest posterior prob } p(C_k|\underline{x})} . p(\underline{x})$$

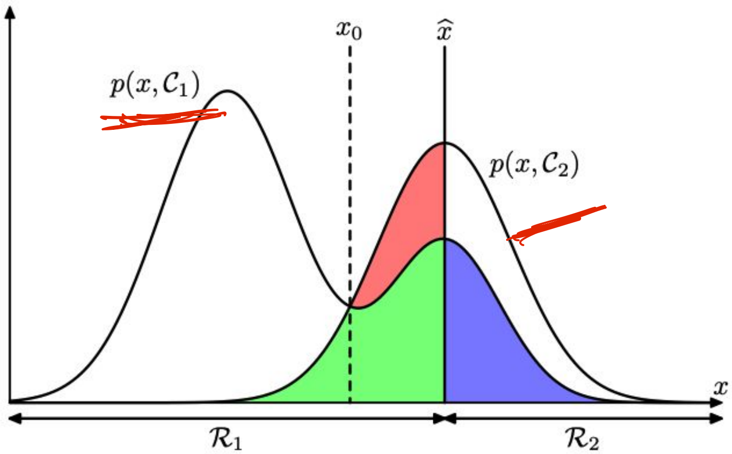
$p(x, C_j)$

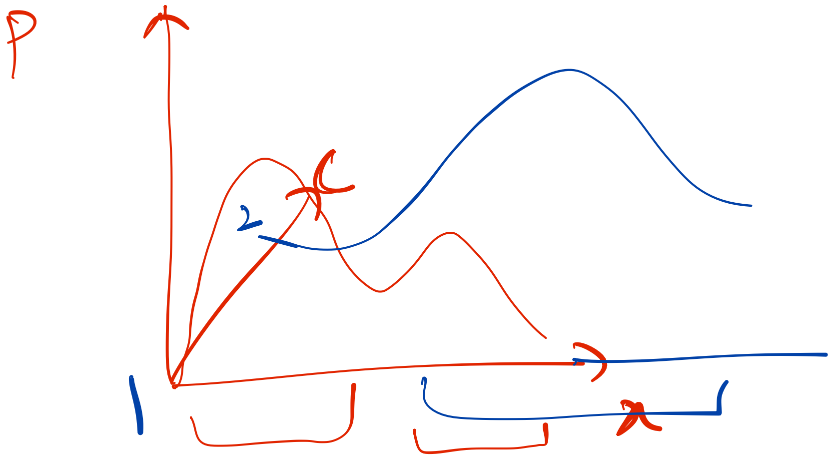
$j \in \{1, \ldots, k\}$

Decision Theory - Misclassification Rate

$k = 1, 2$

$P(\underline{x}, C_k)$





Minimizing the Misclassification Rate - Issues

- Not all errors have the same impact!
- E.g. medical diagnosis
  - E1:
  - E2:

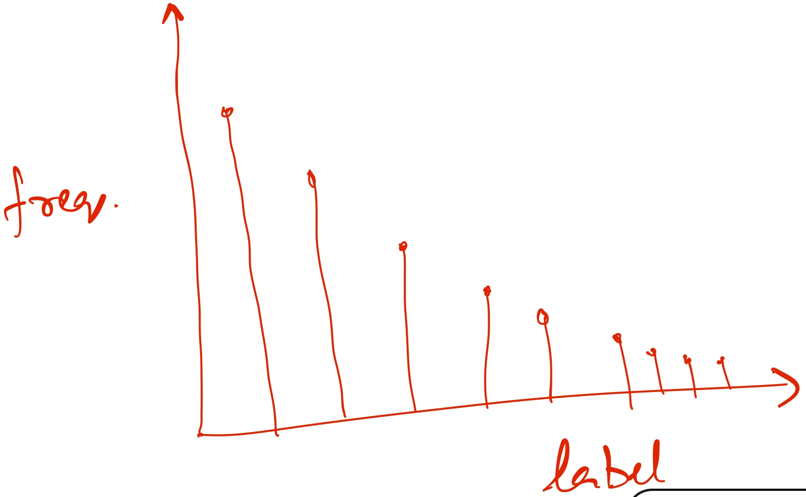
$P(\underline{x}, \hat{t} = D) > P(\underline{x}, \hat{t} = H)$ when $t = H$

$P(\underline{x}, \hat{t} = H) > P(\underline{x}, \hat{t} = D)$ when $t = D$

Minimizing the Misclassification Rate - Issues

- Class imbalance
  - May lead to skewed view of the classifier's performance

1%



Expected Loss

- Possible solution: use different weights for different error types

pred →

$$\mathbf{L} = \begin{array}{cc} C & H \\ \end{array}$$

$$\begin{array}{c|c} & \begin{array}{cc} C & H \end{array} \\ \begin{array}{c} C \\ H \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

↓ G.T.

$$\sum_{k} \sum_{j} \int_{R_j} L_{kj}\, p(\underline{x}, C_k)\, d\underline{x}$$

another way to write

$$\mathbb{E}[L] = \sum_{k,j} L_{k,j} \int_{\mathcal{R}_j} p(x, C_k) dx$$

Minimize the expected loss: (assign x to Ck if) $\sum_{j=1}^{K} L_{kj}\, P(C_k|\underline{x})$ is minimal

Classification Strategies

- Discriminant functions
  - Direct functions of i/p to target $t = y(\mathbf{x}, \mathbf{w})$
- Probabilistic Discriminant models
  - Posterior class probabilities $p(C_k/\mathbf{x})$
- Probabilistic generative models
  - Class-conditional models $p(\mathbf{x}/C_k)$
  - Prior class probabilities $p(C_k)$

Next
Probabilistic Generative Models

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence
& Learning Lab