Bayesian Decision Theory in ML

CAI 5107: Machine Learning

Instructor: Anowarul Kabir

Email: akabir@usf.edu

Fall 2025

Why probabilistic framework?

- A key issue in ML is **uncertainty**
- Why uncertainty:
  - Incomplete or ambiguous data
  - Noisy or wrong information/labeling
- Probability theory and decision theory give us a framework to make optimal decisions or predictions under uncertainty
- Bayesian decision theory: A probabilistic framework for making optimal decisions

Some intuitions

- Data might come from a process that is not completely known
- Example: coin toss
  - We only know the outcome of the toss
- Arguably, if we have extra knowledge of the coin composition, initial position, the force and direction used while coin tossing, and so on, we could predict exact coin toss outcome.
- Uncertainty: observable vs unobservable variables
- Since we do not have access to the data generation process, we model it as random and use probability theory to analyze it

Decision Theory

- Training data: input x and target y

- Inference: Joint probability distribution *P(x, y)*
  - Use training data to learn a model or a hypothesis

- Decision step: Make optimal decision
  - Use the model prediction scores to make optimal class assignments

Decision Theory Goal

- Goal: Minimize the number of mistakes or misclassifications
- We need to define a rule that assigns each input, x, to one of the possible classes, C_k
- Such rules divides the input space into regions R_k s.t. all points in R_k are assigned to C_k

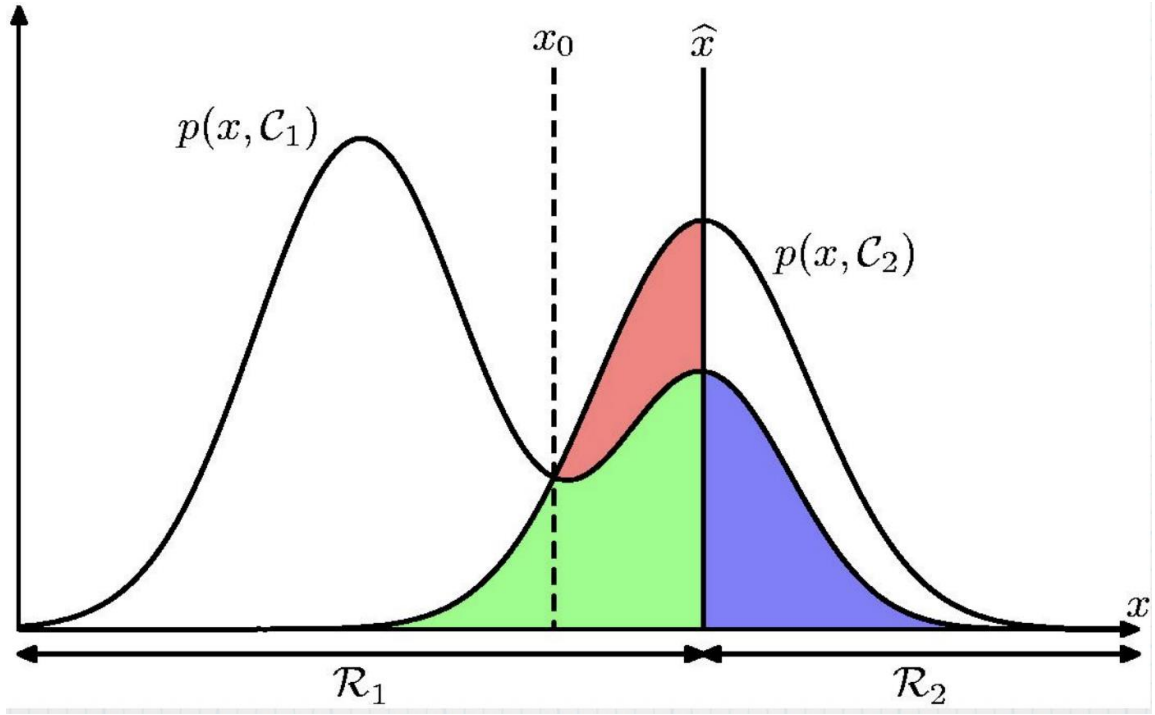
Decision Theory: Binary classification

- Goal: Minimize the number of mistakes
  - Use probabilities in decision making
- Classification: Two classes
  - Classify customers as low- vs. high-risk
- Observable variables: income (X_1), savings (X_2)
- Two classes:
  - $C = 1$ indicating high-risk customers
  - $C = 0$ indicating low-risk customers
- Inference: $P(C|X_1, X_2)$
- Decision: We want to assign a class C for input x s.t. an error measure is minimized

Optimal decision

- Choose C_1 if $P(C_1, x) > P(C_2, x)$
- Choose C_2 if $P(C_2, x) > P(C_1, x)$

- Appling Baye's rule, we can find
  - Choose C_1 if $P(C_1 \mid x) > P(C_2 \mid x)$
  - Choose C_2 if $P(C_2 \mid x) > P(C_1 \mid x)$



Decision Theory

- According to the Bernoulli random variable:
  - $P(C=1) = p$
  - $P(C=0) = q = 1 - p = 1 - P(C=1)$, and
  - $p+q = 1$
- Decision:
  - $C = 1$ *if* $P(C=1 \mid X_1=x_1, X_2=x_2) > 0.5$
  - $C = 0$ overwise
- Mistakes: the probability of error
  - $1 - max(P(C=1|x_1, x_2), P(C=0|x_1, x_2))$

Decision Theory

- Representing the observed variables in a vector format
  - $x = [x_1, x2]^T$
- And using Baye's rule:
  - $P(C, x) = P(C|x)\ P(x) = P(x|C)\ P(C)$
  - $P(C|x) = P(x|C)\ P(C) / P(x)$
- *Prior probability*: $P(C=1)$
  - The proportion of high-risk customers regardless of the observables
  - a.k.a prior knowledge
  - $P(C=1) + P(C=0) = 1$
- *Class likelihood*: $P(x|C)$
  - Conditional probability of an event belong to *C* having the observables *x*

Decision Theory

- *Evidence*: *P(x)*
  - Marginal probability
  - The probability of an observation *x* seen regardless of the classes
  - $P(x) = P(x|C=1)\,P(C=1) + P(x|C=0)\,P(C=0)$
- *Posterior* = likelihood x prior / evidence
  - $P(C=1|x) = P(x|C=1)\,P(C=1) / P(x)$
  - $P(C=0|x) = P(x|C=0)\,P(C=0) / P(x)$

- $P(C=1|x) + P(C=0|x) = 1$

Decision Theory: K-class classification

- For K-class classification: (mutually exclusive classes)
  - $P(C_i) >= 0$
  - $\Sigma_k P(C_k) = 1$
- Using Base's rule:
  - $P(C_i \mid x) = P(x|C_i)\, P(C_i) / P(x)$
  - $P(C_i \mid x) = P(x|C_i)\, P(C_i) / \Sigma_k P(x|C_k)\, P(C_k)$
- Baye's classifier:
  - Choose C_i if $P(C_i|x) = max_k P(C_k \mid x)$

- Do we know the prior ($P(C)$) and the likelihoods ($P(x|C)$)?
  - Often, we need to estimate from the given training samples

Loss of Misclassification

- Minimize the expected loss
- Not all decisions are equally good
- Assume, a_i is the action to assign input x to class C_i and \lambda$_{ik}$ as the incurred loss of taking that action.
- The expected risk:
  - $R(a_i, x) = \Sigma_k \Delta_{ik} P(C_k \mid x)$
- Decision: choose the action with minimum risk
  - Choose a_i if $R(a_i, x) = min_k R(a_i, x)$

Example: For 0/1loss

- For K actions where an action a_i is defined as following
  - a_i: the action of assigning x to C_i
  - where i € {1, K}
- The 0/1 loss is
  - $\Delta_{ik} = 0$ *if i =k, all correct decisions have no loss*
  - $\Delta_{ik} = 1$ *if i != k, all incorrect decisions are equally costly*
- The risk of taking action a_i
  - $R(a_i|x) = \Sigma_k \Delta_{ik} P(C_k \mid x)$
    - *1 - P(Ci|x)*
- To minimize risk, we choose the most probable class

Discriminative Methods

- Solve directly the inference problem of estimating the class posterior probabilities $P(C_k \mid x)$

- Use decision theory to determine class membership for each new input x

Discriminant Functions

- Classification can be thought of learning a set of discriminant functions, g_i s.t.
  - Choose C_i *if* $g_i(x) = max_k\, g_k(x)$
- Maximum discriminant function ~ minimum conditional risk
  - $g_i(x) = P(C_i \mid x) = P(x|C_i)\, P(C_i) / P(x)$
  - $g_i(x) \sim P(x|C_i)\, P(C_i)$
- The discriminant functions divide the feature space into *K* regions

Likelihood ratio

- For binary classfication, the likelihood ratio is defined as
  - $P(x|C_1) / P(x|C_2)$
- What is the discriminant function w.t.o the likelihood ratio?
  - We can choose discriminant function as
    - $G(x) = P(C_1 \mid x) / P(C_2 \mid x)$
    - Choose C_1 if $g(x) > 1$ and C_2 otherwise
  - Apply Baye's rule
  - If the priors are equal, the discriminat is the likelihood ratio

Log odds

- For binary classfication, the log odds is defined as
  - $log\ [P(x|C_1) / P(x|C_2)]$
- What is the discriminant function i.t.o the log odds?
  - We can choose discriminant function as
    - $G(x) = log\ [P(C_1 \mid x) / P(C_2 \mid x)]$
    - Choose C_1 if g(x) > 1 and C_2 otherwise
  - Apply Baye's rule
  - If the priors are equal, the discriminat is the log likelihood ratio

Generative Methods

- Solve the inference problem of estimating the class conditional densities $P(x|C_k)$ for each C
- *Example: Conditional image generation*

Reading materials

- Chapter 3: Bayesian Decision Theory
  - 3.1-3.5