Introduction to ML

September 29th, 2021

Model Selection & Generalization

- Learning is an ill-posed problem; data is not sufficient to find a unique solution
  - Every sample is boolean (d-dimension 2^d possible input function)
  - Every output is Boolean (input-output pair of hypotheses function left 2^{2^d})
  - Each data sample we kills half (the rest are consistent – different function gives the same answer))
    - With N samples there remains $2^{2^d - N}$
    - Impossible for a UNIQUE answer, we usually only see a fraction of data samples

All possible hypothesis functions that takes tuple binary input to binary output

- we have a 2d- binary input (0,1), output is 0 or 1
- there are 2^d possible combination of binary **input** (this is a set with 2^d elements)
- we have another set that is has 2 element (yes/no)
- How many ways we can take 2^d elements to 2 possible outcomes?
    - should be 2^{2^d}

If you have N samples observed, how many possible options are left:

- $2^{2^d - N}$

- take d=2 as an example:
- first possible function (left input, right output)
- (0,0)|0
- (0,1)|0
- (1,0)|0
- (1,1)|0
- second possible function
- (0,0)|1
- (0,1)|0
- (1,0)|0
- (1,1)|0
- third possible function
- (0,0)|0
- (0,1)|1
- (1,0)|0
- (1,1)|0
- fourth possible function
- (0,0)|0
- (0,1)|0
- (1,0)|1
- (1,1)|0
- ... so on, there are total of 2^4 -> 2^2^2 -> 16 functions

Process of of a Supervised Learner

1. Model: $g(x|\theta)$

2. Loss function: $E(\theta|\mathcal{X}) = \sum_t L(r^t, g(x^t|\theta))$

3. Optimization procedure:

$$\theta^* = \arg\min_\theta E(\theta|\mathcal{X})$$

CHAPTER 3:

Bayesian Decision Theory

Classification

Observation → measurable input

- Credit scoring: Inputs are income and savings.

  Output is low-risk vs high-risk

- Input: $\boldsymbol{x} = [x_1, x_2]^T$,Output: C $\in$ {0,1}

- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

ERROR? 1 – max (P(C=1)|x1,x2), P(C=0|x1, x2))

Bernoulli rv
Condition on
x1, x2

Maximum A-posterior decision rule (find the class with highest probability)

Guarantee the error is smallest as measured in terms or probability

Bayes' Rule

What is C (high risk/low risk) given the two inputs

prior *likelihood*

posterior

$$P(C \mid \mathbf{x}) = \frac{P(C)\,p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

evidence

Total probability theorem

$$P(C=0) + P(C=1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} \mid C=1)P(C=1) + p(\mathbf{x} \mid C=0)P(C=0)$$

$$p(C=0 \mid \mathbf{x}) + P(C=1 \mid \mathbf{x}) = 1$$

Bayes' Rule: *K*>2 Classes

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{p(\mathbf{x})}$$
$$= \frac{p(\mathbf{x} \mid C_i) P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k) P(C_k)}$$

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

Make a 'classification' decision by looking at the posterior distribution

Losses and Risks

- Actions: α_i
  - This concept is important in cases when 'loss' associated to each action may not be the same
  - Finance, medical, …so on
- Loss of α_i when the state is C_k : λ_{ik}
- **Expected risk** (Duda and Hart, 1973)

Loss of making a decision to assign input to class i, when the true class is k

Minimizing expected risk by picking that action i

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x})$$

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

Losses and Risks: 0/1 Loss

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

Correct decision: no loss
Incorrect decision : 1 loss

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k \mid \mathbf{x}) \\ &= 1 - P(C_i \mid \mathbf{x}) \end{aligned}$$

All class except that correct one

For minimum risk, choose the most probable class

Losses and Risks: Reject

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

Additional action when in doubt

Loss incurred when in doubt (reject as another action

$$R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$$

Probability sums to 1

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

Risk needs to be minimized

$$\begin{array}{ll} \text{choose } C_i & \text{if } P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x}) \ \forall k \neq i \text{ and } P(C_i \mid \mathbf{x}) > 1 - \lambda \\ \text{reject} & \text{otherwise} \end{array}$$

Has to be greater than reject

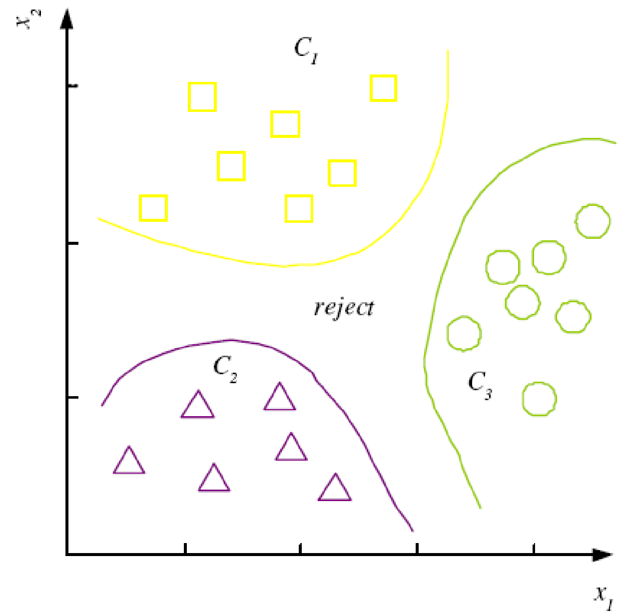
Discriminant Functions

Classification rule: pick one such function that maximizes

choose C_i if $\boxed{g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})}$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i \mid \mathbf{x}) \\ P(C_i \mid \mathbf{x}) \\ p(\mathbf{x} \mid C_i) P(C_i) \end{cases}$$

Three different discriminant functions



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

Divides feature space into K region
For those inputs x, find the function that give the largest value (use that function to carve out a region

K=2 Classes

- Dichotomizer (K=2) vs Polychotomizer (K>2)
- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\text{choose } \begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

- *Log odds:*

A discriminant function

$$\log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})}$$

Needs only 1 value to make a decision

Utility Theory

- Prob of state k given exidence $\boldsymbol{x}$: $P\,(S_k|\boldsymbol{x})$
- Utility of α_i when state is k: U_{ik}
- Expected utility:

Loss: negative utility

Pretty much terminology and used quite often in game theory, economics so on

$$EU(\alpha_i\,|\,\mathbf{x}) = \sum_k U_{ik} P(S_k\,|\,\mathbf{x})$$

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i\,|\,\mathbf{x}) = \max_j EU(\alpha_j\,|\,\mathbf{x})$$

Association Rules

- Association rule: $X \rightarrow Y$
- *People who buy/click/visit/enjoy* ***<u>X</u>*** *are also likely to buy/click/visit/enjoy* ***<u>Y</u>****.*
  - Dependency, it's not cause and effect but good enough for making a business decision

- A rule implies association, not necessarily causation.

Measures for assessing association

Association measures

Joint probability,
Make this large (increase the basis)
Significance of the rule
Useless is this number is low

- Support ($X \to Y$):

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence ($X \to Y$):

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

- conditional probability
- Should be larger than P(Y)
- **Strength** of the association rule

- Lift ($X \to Y$):

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

$$= \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

- If independent, this is 1
- If lift>1 X makes Y more likely

Example

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

SOLUTION:

$$\begin{aligned} \text{milk} \rightarrow \text{bananas} \quad &: \quad \text{Support} = 2/6, \text{Confidence} = 2/4 \\ \text{bananas} \rightarrow \text{milk} \quad &: \quad \text{Support} = 2/6, \text{Confidence} = 2/2 \\ \text{milk} \rightarrow \text{chocolate} \quad &: \quad \text{Support} = 3/6, \text{Confidence} = 3/4 \\ \text{chocolate} \rightarrow \text{milk} \quad &: \quad \text{Support} = 3/6, \text{Confidence} = 3/5 \end{aligned}$$

In making a decision, take ‘high support + high confidence rule”