Machine Learning in Practice: a Crash Course

Lecture 3: Bayesian

胡津铭
DolphinDB

Recap

- Machine Learning Pipeline
- Generalization
- Metrics

Today

- Bayesian Decision Theory
- Naïve Bayes Classifier

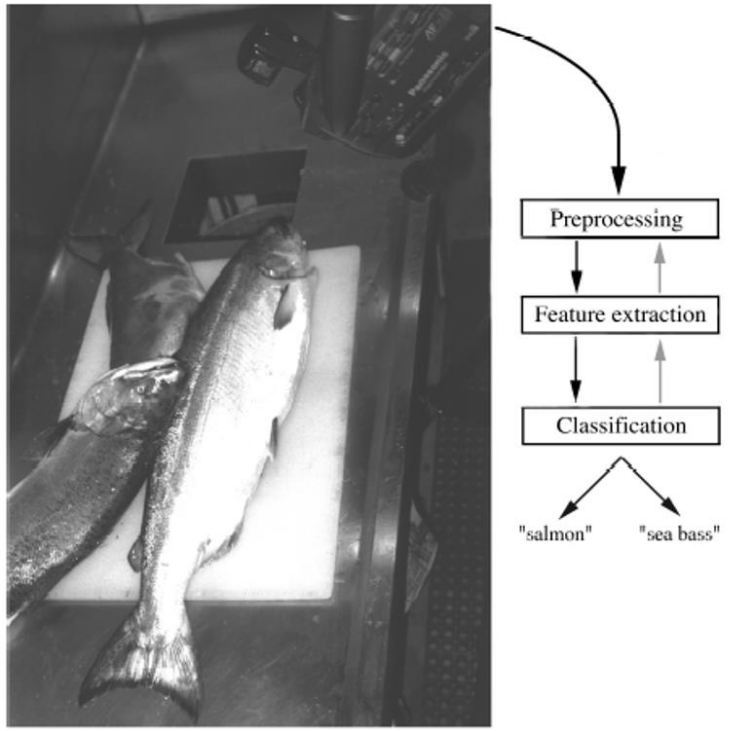
Today

- Since we are going to talk about models in ML
- Big picture of machine learning models:
- Learning = Representation + Evaluation + Optimization

Bayesian Decision Theory

- Decision problem posed in probabilistic terms

- x: sample
- y: state of the nature, often is class label
- $P(y|x)$: given x, what is the probability of the state of the nature
- e.g. $P(y = 1|x)$



Bayes' Theorem

- Conditional probability: $P(A|B) = P(A,B)/\ P(B)$.
- Independence: A and B are said to be independent if and only if P(A, B) = P(A) P(B).

- Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Illustration

A	0	0	1	1	1	0
B	0	1	1	0	1	1

- $P(A = 1) =$, $P(A = 0) =$
- $P(B = 1) =$, $P(B = 0) =$
- $P(A = 1, B = 1) =$
- $P(A = 1|B = 1) =$
- $P(A = 1|B = 1) =$
  - Bayes' Theorem

Prior

lightness	1	2	Total
See bas	5	10	15
salmon	6	5	11

- Reflects our prior knowledge about how likely we are to observe a sea bass or salmon
- The catch of salmon and sea bass is equally probable
  - $P(y_1) = P(y_2)$ (uniform priors)
    - Is this always the case?
    - For example, bomb/cancer detection?
  - $P(y_1) + P(y_2) = 1$ (exclusivity and exhaustivity)
- Decision rule with only the prior information
  - Decide y_1 if $P(y_1) > P(y_2)$, otherwise y_2
  - What's the problem with this? It does not utilize any feature.

Likelihood

lightness	1	2	Total
See bas	5	10	15
salmon	6	5	11

- In sea bass/salman example
- $P(x|y_1)$ and $P(x|y_2)$ describe the difference in lightness feature between populations of sea bass and salmon
- $P(x|y_j)$ is called the **likelihood** of y_j with respect to x;
- the category y_j for which $P(x|y_j)$ is large is more likely to be the true category
- **Maximum likelihood decision**
  - Decide y_1 if $P(x|y_1) > P(x|y_2)$ otherwise y_2

Posterior

lightness	1	2	Total
See bas	5	10	15
salmon	6	5	11

- Bayes formula

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

$$P(x) = \sum_{i=1}^{k} P(x|y_i)P(y_i)$$

Posterior = (**Likelihood** × **Prior**) / Evidence

- Evidence $P(\boldsymbol{x})$ can be viewed as a scale factor that guarantees that the posterior probabilities sum to 1
- **Posterior** ∝ **Likelihood** × **Prior**

Optimal Bayes Decision Rule

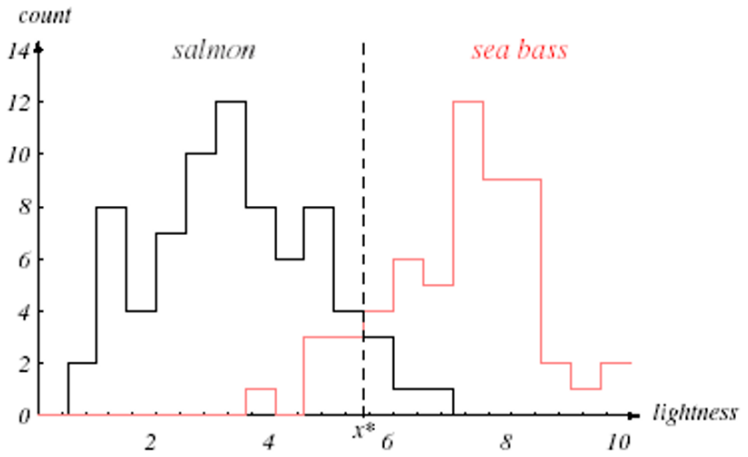
- Decide y_1 if $P(y_1|x) > P(y_2|x)$, otherwise y_2
- Bayes decision rule minimizes the probability of error, that is the term Optimal comes from.
  - But why?

Bayesian Decision Theory – Generalization

- Generalization of the preceding ideas
- Use of more than one feature (p features)
- Use of more than two classes (c classes)
- Are errors equal?

Bayesian Risk

- $E_{ij} = E(\widehat{y_i}|\ y_j)$
- $R(\widehat{y_1}|x) = E_{11}P(y_1|x) + E_{12}P(y_2|x) = E_{12}P(y_2|x)$
- $R(\widehat{y_2}|x) = E_{21}P(y_1|x) + E_{22}P(y_2|x) = E_{21}P(y_1|x)$
- Decide y_1 if $R(\widehat{y_1}|x) < R(\widehat{y_2}|x)$, otherwise y_2
- Special Case, if $E_{12} = E_{21}$, then this is?
  - Posterior decision rule!
- Recall the sea bass/salmon, what is it?



Example 1: Two-class classification

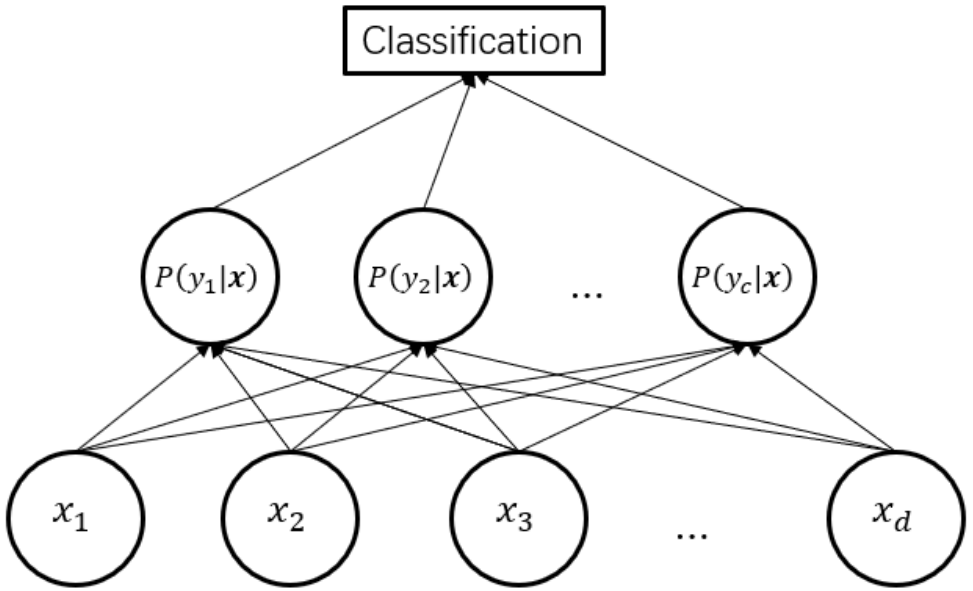
- Decide y_1 if $R(\widehat{y_1}|x) < R(\widehat{y_2}|x)$, otherwise y_2
- $y_1 \; if \; E_{21}P(y_1|x) > E_{12}P(y_2|x), otherwise \; y_2$
- $E_{21} \; P(x|y_1) \; P(y_1) > E_{12}P(x|y_2) \; P(y_2)$
- The preceding rule is equivalent to the following rule:
- If $\frac{P(\boldsymbol{x}|y_1)}{P(\boldsymbol{x}|y_2)} > \frac{E_{12}}{E_{21}} \times \frac{P(y_2)}{P(y_1)}$
- "If the likelihood ratio exceeds a threshold value that is independent of the input pattern x, we can take optimal actions"

Example 2: Multi-class classification

- If we predict as $\hat{y}_i$ and the label is y_j then:
- the decision is correct if $i = j$ and in error if $i \neq j$
- Seek a decision rule that minimizes the **probability of error** or the **error rate** or maximize the **accuracy**
- $E(\hat{y}_i|y_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$
- Minimizing the risk ➔ Maximizing the posterior $P(y_i|\boldsymbol{x})$

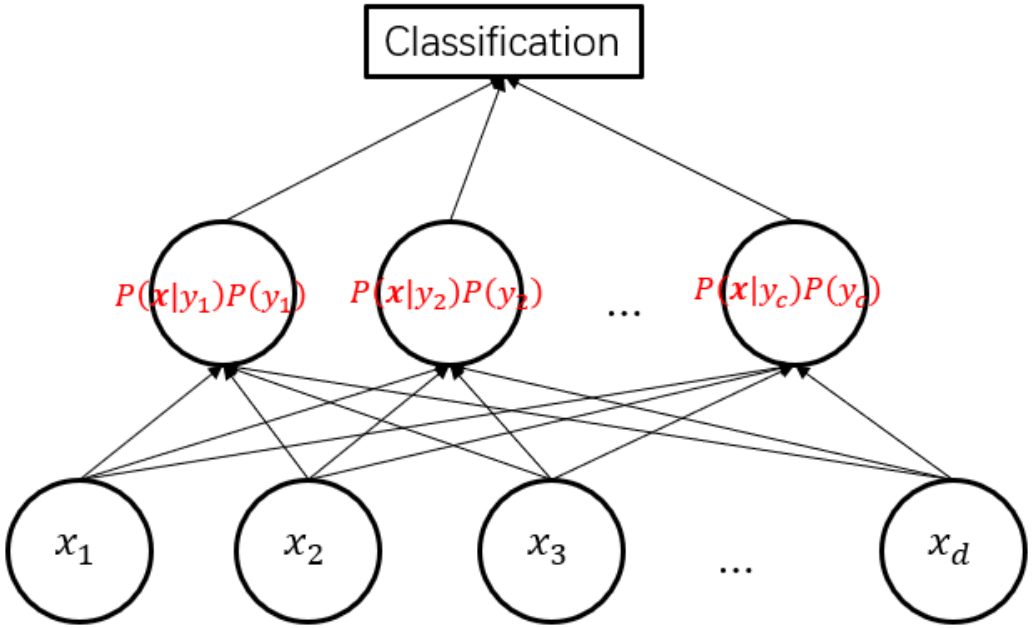
Bayesian Decision Rule

- Decide $\boldsymbol{y_i}$ if $P\ (\boldsymbol{y_i}|\ x)\ >\ P(\boldsymbol{y_j}|\ x)\ \forall j \neq i$



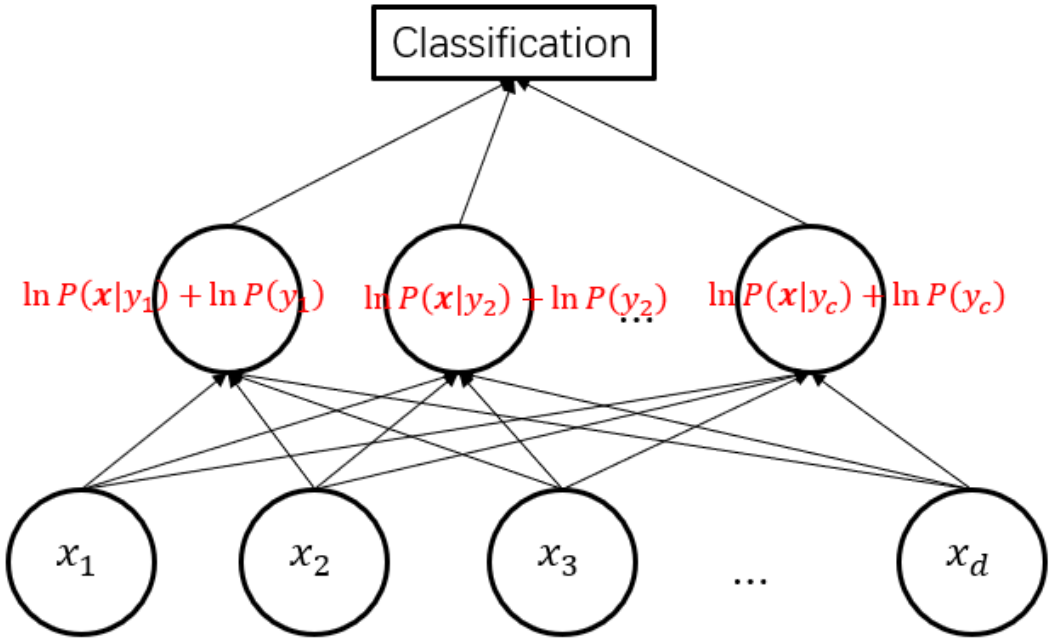
Bayesian Decision Rule

- Decide $\boldsymbol{y_i}$ if $P\ (\boldsymbol{y_i}|\ x)\ >\ P(\boldsymbol{y_j}|\ x)\ \forall j \neq i$

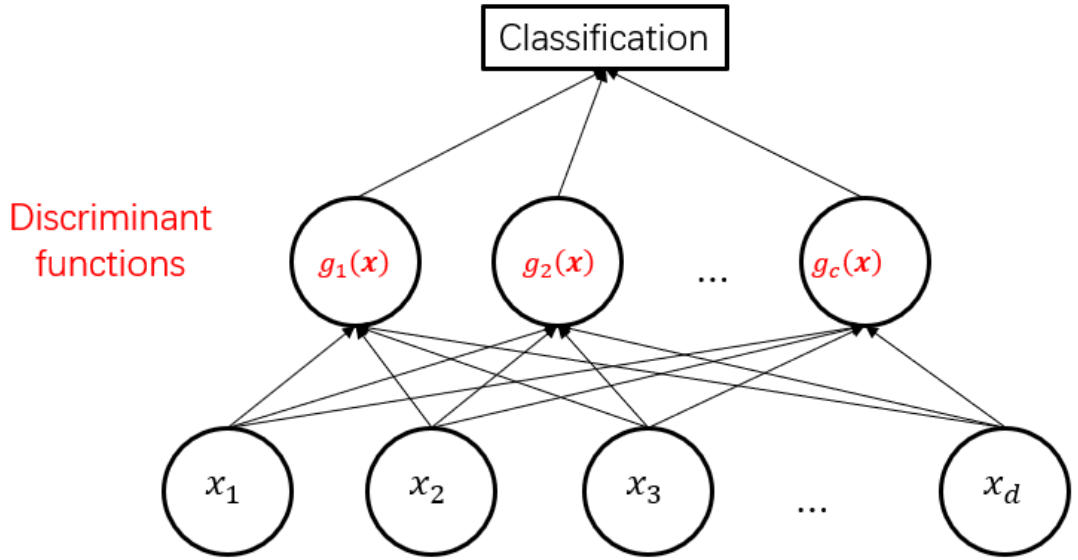


Bayesian Decision Rule

- Decide $\boldsymbol{y_i}$ if $P\ (\boldsymbol{y_i}|\ x)\ >\ P(\boldsymbol{y_j}|\ x)\ \forall j \neq i$



Discriminant Functions and Classifiers



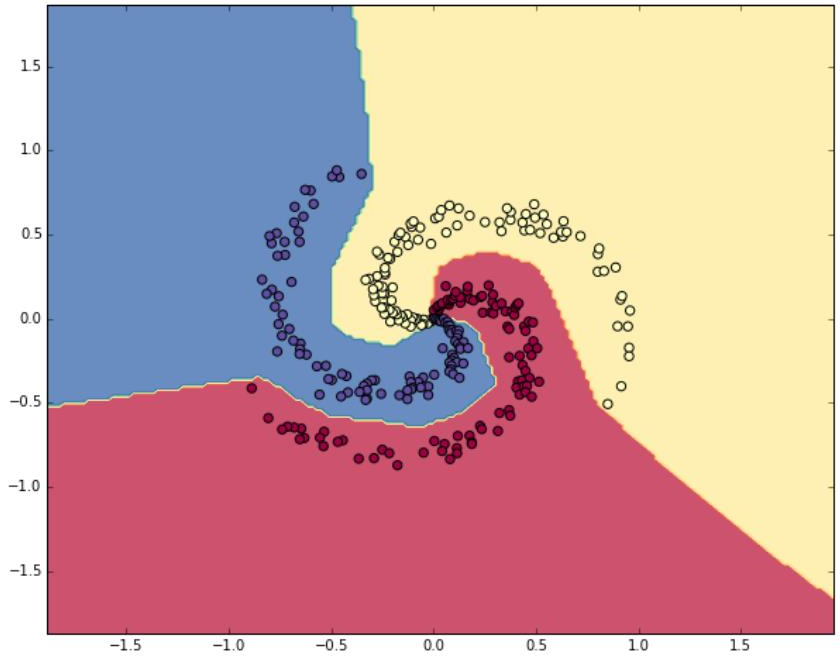
- Set of discriminant functions: $g_i(\boldsymbol{x}),\ i = 1, \cdots, c$
- Classifier assigns a feature vector $\boldsymbol{x}$ to class y_i if:

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \qquad \forall j \neq i$$

Decision Regions and Surfaces

- Effect of any decision rule is to divide the feature space into c decision regions
- If $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})\ \forall j \neq i$, then $\boldsymbol{x} \in \boldsymbol{\mathcal{R}}_i$

(Region $\boldsymbol{\mathcal{R}}_i$ means assign $\boldsymbol{x}$ to y_i)



So Far…

- Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(y_i)$: priors
    - $P(x \mid y_i)$: class-conditional densities

    Unfortunately, we rarely have this complete information!

- Design a classifier based on a set of labeled training samples (supervised learning)
  - Estimate it from the data

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(y_k) = \frac{N_{y_k}}{N}$
  - e.g., $P(No) = 7/10, P(Yes) = 3/10$
- For discrete attributes:

$$P(x_i|y_k) = \frac{|x_{ik}|}{N_{y_k}}$$

  - where $|x_{ik}|$ is number of instances having attribute x_i and belongs to class y_k
  - Examples:

$P(Status = Married|No) = 4/7$

$P(Refund = Yes|Yes) = 0$

How to Estimate Probabilities from Data?

- For continuous attributes:
  - **Discretize** the range into bins
    - one ordinal attribute per bin
  - **Two-way split:** $(x < v)$ or $(x > v)$
    - choose only one of the two splits as new attribute
  - **Probability density estimation:**
    - Assume attribute follows a normal distribution or some other distribution

Bayesian Decision Rule Notebook Demo

So Far…

- Design a classifier based on a set of labeled training samples (supervised learning)
  - Estimate the probability from the data
  - Need sufficient no. of training samples for estimating class-conditional densities, especially when the dimensionality of the feature space is large
  - Curse of Dimensionality

Naïve Bayes Classifier

- Given $\boldsymbol{x} = (x_1, \cdots x_p)$
  - Goal is to predict class y
  - Specifically, we want to find the value of y that maximizes $P(y|\boldsymbol{x}) = P(y|x_1, \cdots x_p)$
  - $P(y|x_1, \cdots x_p) \propto P(x_1, \cdots x_p \mid y)P(y)$

- Conditional independence assumption among features
  - $P\big(x_1, \cdots x_p|y\big) = P(x_1|y) \cdots P\big(x_p|y\big)$
  - Clearly not true for most applications, thus the name “Naïve”

Naïve Bayes Classifier

- Conditional independence assumption among features
  - $P(x_1, \cdots x_p|y) = P(x_1|y) \cdots P(x_p|y)$
  - What are the potential problems?
    - Float point underflow: take a log!
    - One zero will zero-out all
  - Smoothing

$$P(x_i|y_k) = \frac{|x_{ik}|+1}{N_{y_k}+K}$$

Naïve Bayes Classifier Notebook Demo

Naïve Bayes Summary

- Advantages
  - Robust to isolated noise points
  - Handle missing values by ignoring the instance during probability estimate calculations
  - Robust to irrelevant attributes
- Disadvantages
  - Independence assumption may not hold for some attributes
- But it works for spam/ham classifier very well, why?
  - Many words are sort of independent
  - $P(x_1, \cdots x_p|y) = P(x_1|y) \cdots P(x_p|y)$
  - Even if the value may not be equal, the order may keep
  - Very efficient to train/predict even with a large vocabulary, i.e. a lot of features

Question?

Thanks and welcome to give us suggestions and feedbacks afterwards.