Parametric density estimation

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 2

Bayesian decision theory (cont'd)

loss and risk

- Define
  - action α_i as the decision to assign the input to class C_i
  - λ_{ik} as the loss incurred for taking α_i when the input actually belongs to C_k (if we allow abuse of notation, we can say $\boldsymbol{x} \in C_k$).
- Then the **expected risk** for taking α_i is

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

loss and risk

- $R(\alpha_i|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\boldsymbol{x})$
- In the special case of **0/1 loss**, where $\lambda_{ik} = \begin{cases} 0 & \text{if} \quad i = k \\ 1 & \text{if} \quad i \neq k \end{cases}$
- $R(\alpha_i|\boldsymbol{x}) = \sum_{k \neq i} P(C_k|\boldsymbol{x}) = 1 - P(C_i|\boldsymbol{x})$

reject

- In the above, we already have actions α_i as the decision to assign the input to class $C_i,\ \ i = 1,2,\cdots,K$
- Let's define an additional action of **reject** (not making any decision, indecisive): α_{K+1}
- By modifying the 0/1 loss, a possible loss function is

$$\lambda_{ik} = \begin{cases} 0 & \text{if} \quad i = k \\ 1 & \text{if} \quad i \in [K] - \{k\} \\ \lambda & \text{if} \quad i = K+1 \end{cases} = \begin{cases} 0 & \text{if} \quad i = k \\ \lambda & \text{if} \quad i = K+1 \\ 1 & \text{otherwise} \end{cases}$$

reject

$$\lambda \sum_{k=1}^{K} P(C_k|x) = \lambda \cdot 1$$

$\qquad \qquad \qquad \qquad \qquad \qquad \parallel \qquad \qquad \qquad \qquad \parallel$

- The risk of reject is $R(\alpha_{K+1}|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda P(C_k|\boldsymbol{x}) = \lambda$
- The risk of choosing C_i is $1 - P(C_i|\boldsymbol{x})$

reject

- The optimal decision rule:

  - Choose C_i if

    (1) $R(\alpha_i|\boldsymbol{x}) < R(\alpha_k|\boldsymbol{x})$ for all $k \neq i$ and

    (2) $R(\alpha_i|\boldsymbol{x}) < R(\alpha_{K+1}|\boldsymbol{x})$

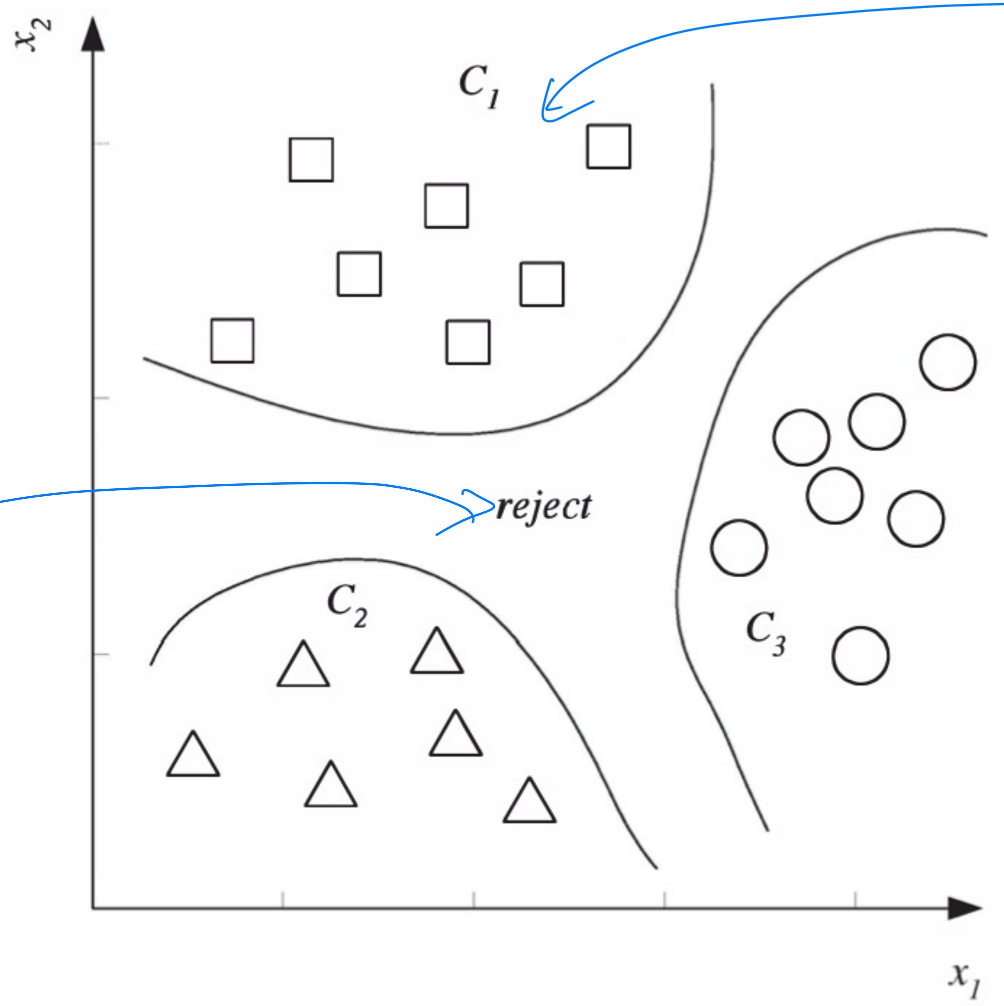
  - Reject if

    $R(\alpha_{K+1}|\boldsymbol{x}) < R(\alpha_i|\boldsymbol{x})$ for all i

reject

- The optimal decision rule:
  - Choose C_i if
    (1) $P(C_i|\boldsymbol{x}) > P(C_k|\boldsymbol{x})$ for all $k \neq i$ and
    (2) $P(C_i|\boldsymbol{x}) > 1 - \lambda$
  - Reject if
    $P(C_i|\boldsymbol{x}) < 1 - \lambda$ for all i

decision region and decision boundary



$P(C_1 | x_1, x_2) > 1-\lambda$ and

$P(C_1 | x_1, x_2) > P(C_j | x_1, x_2)$

for $j = 2, 3$.

$1-\lambda > P(C_j | x_1, x_2)$

for $j = 1, 2, 3$.

image taken from [AI]

discriminant functions

- Classification can be viewed as implementing a set of **discriminant functions** $g_i(\boldsymbol{x}), i = 1, \cdots, K,$ such that we

$$\text{choose } C_i \text{ if } g_i(\boldsymbol{x}) = \max_{k=1,\cdots,K} g_k(\boldsymbol{x})$$

- We can choose $g_i(\boldsymbol{x}) = -R(\alpha_i|\boldsymbol{x})$ or choose it to be $P(C_i|\boldsymbol{x})$

- We can also put $g_i(\boldsymbol{x}) = p(\boldsymbol{x}|C_i)P(C_i)$

because $P(C_i|x) = \dfrac{p(x|C_i)\,P(C_i)}{p(x)}$

Same for all classes

maximum likelihood estimator

parametric approach

- In a parametric method
  - A sample is drawn from some distribution that obeys a known model.
  - This model is defined up to a small number of parameters.
  - e.g. $\mathcal{N}(\mu, \sigma^2)$ is a parametric model that depends on two parameters: μ and σ.

statistic

- A **statistic** is any value that is calculated from a given sample.

- A statistic is said to be **sufficient** (for the underlying parametric model) if:
  - no further information can be inferred from other statistics calculated from the same sample

statistic

independent, identically distributed

- e.g. $x_1, x_2, \cdots, x_N \sim^{i.i.d.} \mathcal{N}(\mu, \sigma^2)$
  - the sample mean $m = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - the sample variance $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$
  - then (m, s^2) is a **sufficient statistic** for (μ, σ^2)

- Using sufficient statistics, we can get statistical models with only few parameters.

parametric approach

- We start the study of parametric approaches with the problem of density estimation:

  - $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$, where $\boldsymbol{x}_n \sim^{i.i.d.} p(\boldsymbol{x}|\boldsymbol{\theta})$

  - Want $\boldsymbol{\theta}$ such that $\boldsymbol{x}_n$ is sampled from $p(\boldsymbol{x}|\boldsymbol{\theta})$ <u>as likely as possible</u>.

likelihood

$$p(x_1, x_2, \cdots, x_N | \theta) \overset{\text{by iiid}}{=\!=} p(x_1|\theta)\, p(x_2|\theta) \cdots p(x_N|\theta)$$

- maximize the **likelihood** of θ given $\mathcal{X}$:

$$l(\boldsymbol{\theta}|\mathcal{X}) := p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

- equivalently, maximize the **log-likelihood** of θ given $\mathcal{X}$:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) = \log l(\boldsymbol{\theta}|\mathcal{X}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

Remark: notations here follow [AI]

Bernoulli density

$$X = \begin{cases} 1 & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases}$$

$$\mathbb{P}(X = x) = \underbrace{\boxed{q^{x}(1-q)^{1-x}}}_{p(x|q)}, \quad x \in \{0, 1\}$$

- Parameter: q

Now, if we are given a sample $\mathcal{X} = \{x_n\}_{n=1}^{N}$

By definition, the likelihood is

$$l(q|\mathcal{X}) = p(\mathcal{X}|q) = \prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}$$

The log-likelihood is

$$\mathcal{L}(q|\mathcal{X}) = \sum_{n=1}^{N} x_n \log q + (1-x_n)\log(1-q)$$

Bernoulli density

To maximize $\mathcal{L}(q \mid X)$, we need to solve

$$\max_{q} \underbrace{\sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)}_{f(q)}$$

Setting $\dfrac{d f(q)}{d q} = \sum_{n=1}^{N} \dfrac{x_n}{q} - \dfrac{1 - x_n}{1 - q} = 0$

That gives $\dfrac{\sum_{n=1}^{N} x_n}{q} = \dfrac{N - \sum_{n=1}^{N} x_n}{1 - q}$

That is $q = \dfrac{\sum_{n=1}^{N} x_n}{N}$. We conclude $\hat{q}_{MLE} = \dfrac{\sum_{n=1}^{N} x_n}{N}$.

multinomial density

K states $\{1, 2, \cdots, K\}$.

X takes State i with probability q_i. $\sum_{i=1}^{K} q_i = 1$

Define $X_i = \begin{cases} 1 & \text{if State } i \text{ is taken} \\ 0 & \text{otherwise} \end{cases}$

We can represent X as a vector $(X_1 \ X_2 \ \cdots \ X_K)^T$

one-hot vector

$= (0 \ 0 \ \cdots \ 0 \ 1 \ 0 \cdots 0)^T$

i-th entry if State i is taken

multinomial density

$(x_1, x_2, \cdots, x_K)^T$

$$\mathbb{P}(X = x) = q_1^{x_1} \, q_2^{x_2} \cdots q_K^{x_K} \leftarrow p(x \mid q_1, \cdots, q_K)$$

- parameters: $q_1, q_2, \cdots, q_K$

Given a sample $\mathcal{X} = \{\underline{x_n}\}_{n=1}^{N}$

each x_n is a K-dim one-hot vector.

$$l(q_1, q_2, \cdots, q_K \mid \mathcal{X}) = \prod_{n=1}^{N} \prod_{i=1}^{K} q_i^{x_{ni}}$$

$$\mathcal{L}(q_1, q_2, \cdots, q_K \mid \mathcal{X}) = \sum_{n=1}^{N} \sum_{i=1}^{K} x_{ni} \log q_i$$

multinomial density

To maximize the (log) likelihood, we need to solve

$$\max_{q_1, \cdots, q_K} \sum_{n=1}^{N} \sum_{i=1}^{K} x_{ni} \log q_i$$

$$\text{s.t.} \quad \sum_{i=1}^{K} q_i = 1.$$

multinomial density

Define $L = \sum_{n=1}^{N} \sum_{i=1}^{K} x_{ni} \log q_i - \lambda \left(\sum_{i=1}^{K} q_i - 1 \right)$

Setting $\begin{cases} \dfrac{\partial L}{\partial q_i} = \dfrac{\sum_{n=1}^{N} x_{ni}}{q_i} - \lambda = 0 \\[2ex] \dfrac{\partial L}{\partial \lambda} = \sum_{i=1}^{K} q_i - 1 = 0 \end{cases}$

yields $q_i = \dfrac{\sum_{n=1}^{N} x_{ni}}{\lambda}$. Plugging this in

multinomial density

$$\sum_{i=1}^{K} \frac{\sum_{n=1}^{N} x_{ni}}{\lambda} = 1 .$$

That is, $\lambda = \sum_{n=1}^{N} \overbrace{\left(\sum_{i=1}^{K} x_{ni} \right)}^{1} = N$

Therefore,

$$q_i = \frac{\sum_{n=1}^{N} x_{ni}}{N} .$$

This gives you $\hat{q}_{MLE}$

Questions?

Reference

- *Bayesian decision theory:*
  - *[Al] Ch.3.1-3.4*
  - *[HaTF] Ch.2.4*
- *Maximum likelihood:*
  - *[Al] Ch.4.1-4.3*
  - *[Bi] Ch.2.4*
  - *[HaTF] Ch.2.6, 8.2.2*