Statistic

for machine learning

Tran Trong Khiem

AI lab tranning

2024/05/29

1. Bayesian decision theory
2. Choosing the “right” model
3. Frequentist decision theory
4. Frequentist hypothesis testing

Basics

- **Agent**, has a set of possible actions,$\mathcal{A}$ to choose.
- Actions has **costs** and **benefits**, depend on the **state of nature**, $H \in \mathcal{H}$
  - **loss function** : $l(h, a)$
- **Posterior expected loss** : $\rho(a|x) \hat{=} E_{p(h|x)}[l(h, a)] = \sum_{h \in \mathcal{H}} l(h, a) p(h|x)$
- **optimal policy** $\pi^*(x)$ (Bayes estimator):
  - $\pi^*(x) = arg\min_{a \in \mathcal{A}} \rho(a|x)$
- Classification problems
  - The loss function is $l(y^*, \hat{y})$, should choose label $\hat{y} = 0$ iff:

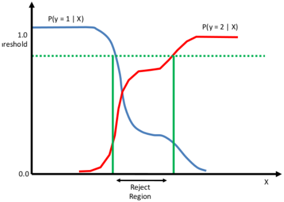
$$l_{00}p_0 + l_{10}p_1 < l_{11}p_1 + l_{01}p_0$$

Classification with the "reject" option

$$\ell(y^*, a) = \begin{cases} 0 & \text{if } y^* = a \text{ and } a \in \{1, \dots, C\}, \\ \lambda_r & \text{if } a = 0, \\ \lambda_e & \text{otherwise,} \end{cases}$$

Where :

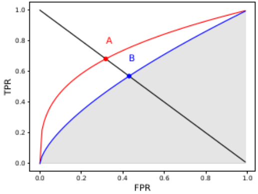
- λ_r is the cost of rejection.
- λ_e is the cost of a classification error.
- **optimal action** is to **pick the reject** action iff :
  - if the **most probable class** has a **probability** below $\lambda^* = 1 - \frac{\lambda_r}{\lambda_e}$



ROC curves

Ideal : Instead of **picking a single threshold**, using a **set of different thresholds**, to **comparing performance**.

- true positive rate (TPR): $TPR_\tau = p(\hat{y} = 1|y = 1, \tau) = \frac{TP_\tau}{TP_\tau + FN_\tau}$
- false positive rate (FPR): $FPR_\tau = p(\hat{y} = 1|y = 0, \tau) = \frac{FP_\tau}{FP_\tau + TN_\tau}$
- **plot TPR vs FPR** as an implicit function of τ



- using the **area under the curve** or **AUC**, AUC scores are better.
- **equal error rate** or **EER**, defined as the **value satisfies** : $FPR = FNR = 1 - TPR$, lower **EER** scores are better.

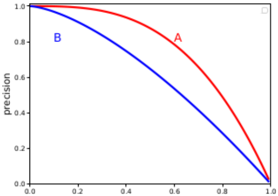
Precision-recall curves

Problem: in imbalance class

- The **ROC curve** is **unaffected** by **class imbalance**.
- **usefulness** of an ROC curve may be **reduce**

Precision-recall curves

- using when **imbalance class** in **negative**.
- **Ideal** : **replace** the FPR by **precision**
- **precision** : $\mathcal{P}(\tau) = p(y = 1|\hat{y} = 1, \tau) = \frac{TP_\tau}{TP_\tau + FP_\tau}$
- **recall** : $\mathcal{R}(\tau) = p(\hat{y} = 1|y = 1, \tau) = \frac{TP_\tau}{TP_\tau + FN_\tau}$



F-scores

- $\frac{1}{F_\beta} = \frac{1}{1+\beta^2}\frac{1}{\mathcal{P}} + \frac{\beta^2}{1+\beta^2}\frac{1}{\mathcal{R}}$
- $\mathcal{R}$ is recall at fixed τ, $\mathcal{P}$ is precision, we set $\beta = 1$, we get :

$$\frac{1}{F_1} = \frac{1}{2}(\frac{1}{\mathcal{P}} + \frac{1}{\mathcal{R}})$$

- F_1 score weights precision and recall equally.$\beta = 2$, if recall is more important.

Regression problems

L2 loss:

- defined as : $l_2(h, a) = (h - a)^2$
- the risk is given by :

$$\rho(a|x) = \mathbb{E}\left[(h-a)^2 \mid x\right] \qquad = \mathbb{E}\left[h^2 \mid x\right] - 2a\mathbb{E}\left[h \mid x\right] + a^2$$

- optimal option: $\frac{\partial \rho(a|x)}{\partial a} = 0$

$$\pi(a) = \mathbb{E}[h|a] = \int hp(h|x)dh$$

L1 loss

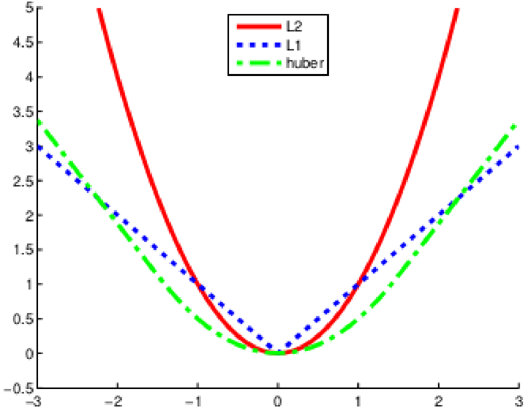
- definded as : $l_1(h, a) = |h - a|$

Huber loss

$$\delta(h, a) = \begin{cases} \frac{r^2}{2} & \text{if } |r| \leq \delta, \\ \delta|r| - \frac{\delta^2}{2} & \text{if } |r| > \delta, \end{cases}$$

where $r = h - a$.

Regression problems



Probabilistic prediction problems

KL, cross-entropy and log-loss

- KL defined as :

$$\mathrm{DKL}(p \parallel q) = \sum_{y \in Y} p(y) \log p(y) - \sum_{y \in Y} p(y) \log q(y)$$

$$= -H(p) + H_{ce}(p, q)$$

- $H(p)$ term is known as the **entropy**
  - measure of uncertainty or variance of p
- $H_{ce}(p, q)$ is the cross-entropy.
- **Optimal:** $q^*(Y \mid x) = \arg\min_q H_{\text{ce}}(q(Y \mid x), p(Y \mid x))$
  - if **true state of nature** is **one hot** distribution:

$$H_{\text{ce}}(\delta(Y = c), q) = -\sum_{y \in Y} \delta(y = c) \log q(y) = -\log q(c)$$

1. Bayesian decision theory
2. Choosing the “right” model
3. Frequentist decision theory
4. Frequentist hypothesis testing

Bayesian hypothesis testing

Problem:

- there are **several candidate model**.
- how to chose the **"right"** model.

Bayesian hypothesis testing:

- There are two models: M_0 (null hypothesis), M_1 (alternative hypothesis)
- If use **0-1 loss**, chose M_1 iff $p(M_1|\mathcal{D}) > p(M_0|\mathcal{D})$
- If using **uniform prior**, $p(M_1) = p(M_0) = 0.5$. Select M_1 iff

$$p(\mathcal{D}|M_1) > p(\mathcal{D}|M_0)$$

Example: Testing if a coin is fair

- model M_0:$p(\mathcal{D}|M_0) = (\frac{1}{2})^N$, where N is the number of coin tosses.
- model M_1:$p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1+N_1,\alpha_0+N_0)}{B(\alpha_1,\alpha_0)}$

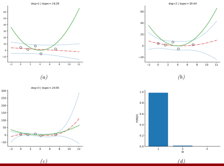
Bayesian model selection

Set $\mathcal{M}$ of more than 2 models, $m \in \mathcal{M}$.

- If using **0-1 loss**, the optimal action is to pick the **most probable model**: $\hat{m} = arg\max_{m \in \mathcal{M}} p(m|\mathcal{D})$
  - $p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in M} p(D|m)p(m)}$, is the posterior over model.
- If the **prior over models** is **uniform**, $p(m) = \frac{1}{|\mathcal{M}|}$, then the MAP model is given: $\hat{m} = arg\max_{m \in \mathcal{D}} p(m|\mathcal{D})$
  - $p(m|\mathcal{D})$ is given by :

$$p(m|\mathcal{D}) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta$$

Example: polynomial regression



Occam's razor

Consider two models,

- simple one m_1, complex one m_2, $p(\mathcal{D}|\hat{\theta}_1, m_1)$ and $p(\mathcal{D}|\hat{\theta}_2, m_2)$ are both large.
- should prefer m_1 , since it is **simpler**.

The complex model :

- put **less prior probability** on the "good" parameters that **explain the data**.
- take **averages in parts of parameter space** with low likelihood.

The simpler model:

- has **fewer parameters**
- prior is **concentrated over a smaller volume**

Conservation of probability mass principle

- Complex model must **spread their predicted probability mass thinly**, not obtain a large probability for any given data.

Connection between cross validation and marginal likelihood

Goal: turn out **marginal likelihood** is **closely related** to the **leave-one-out cross-validation**.

$$p(D|m) = \prod_{n=1}^{N} p(y_n|y_{1:n-1}, x_{1:N}, m) = \prod_{n=1}^{N} p(y_n|x_n, D_{n-1}, m)$$

Where :$p(y_n|x_n, D_{n-1}, m) = \int p(y_n|x_n, \theta)p(\theta|D_{n-1}, m)d\theta$
Use a **plugin approximation** we get :

$$p(y|x, D_{1:n-1}, m) \approx \int p(y|x, \theta)\delta(\theta - \hat{\theta}_m(D_{1:n-1}))\, d\theta = p(y|x, \hat{\theta}_m(D_{1:n-1}))$$

The we get :

$$\log p(D|m) \approx \sum_{n=1}^{N} \log p(y_n|x_n, \hat{\theta}_m(D_{1:n-1}))$$

Information criteria

Problem:

- Bayesian model selection using $p(\mathcal{D}|m) = \int p(\mathcal{D}|m, \theta)p(\theta)d\theta$
- can be **difficult to compute**

Solution: otherway **information criteria**
The Bayesian information criterion (BIC)

- Simple **approximation** to the **log marginal likelihood**

$$\log p(D|m) \approx \log p(D|\hat{\theta}_{\text{MAP}}) + \log p(\hat{\theta}_{\text{MAP}}) - \frac{1}{2}\log|H|$$

- If using **uniform prior**, replace MAP by MLE $\hat{\theta}$:

$$\log p(D|m) \approx \log p(D|\hat{\theta}) - \frac{1}{2}\log|H|$$

Where H is the **Hessian of the negative log joint**, $-log(\mathcal{D}, \theta)$

$$J_{\text{BIC}}(m) = \log p(D|m) \approx \log p(D|\hat{\theta}, m) - \frac{D_m \log N}{2}$$

The Bayesian information criterion (BIC)

Akaike information criterion :

- $L_{\text{AIC}}(m) = -2\log p(D|\hat{\theta}, m) + 2D,$
  - $D = dim(\theta)$

Minimum description length (MDL):

- $C(m) = -log(p(m))$
- $\mathcal{L}_{\text{MDL}}(m) = -\log p(D|\hat{\theta}, m) + C(m)$

Posterior inference over effect sizes and Bayesian significance testing

Problem

- **hypothesis testing** is using $\frac{p(\mathcal{D}|M_0)}{p(\mathcal{D}|M_1)}$
- **computationally** difficult

Bayesian t-test for difference in means

- We have : model m_1 and m_2, N is dataset size, e_i^m is error of m at text i.
- $d_i = e_i^1 - e_i^2$, assume $d_i \sim \mathcal{N}(\Delta, \sigma^2)$, $d = (d_1, ...d_N)$
- If using an **uninformative prior** for (Δ, σ):

$$p(\Delta|d) = \mathcal{T}_{N-1}\left(\Delta \middle| \mu, \frac{s^2}{N}\right)$$

Where $\mu = \frac{1}{N}\sum_{i=1}^{N} d_i$, $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \mu)^2$

Bayesian χ^2-test for difference in rates

Problem : **two model** which are evaluated on **different test sets**.
We have :

- y_m is number of correct examples,N_m trails, **accuracy rate** $\frac{y_m}{N_m}$
  - Assume $y_m \sim \text{Bin}(N_m, \theta_m)$, $\Delta = \theta_1 - \theta_2$, $D = (y_1, N_1, y_2, N_2)$
  - $p(\theta_1, \theta_2|D) = \text{Beta}(\theta_1|y_1 + 1, N_1 - y_1 + 1)\text{Beta}(\theta_2|y_2 + 1, N_2 - y_2 + 1)$

    The posterior for Δ is given by

$$p(\Delta|D) = \int_0^1 \int_0^1 \mathbb{I}(\Delta = \theta_1 - \theta_2)\, p(\theta_1|D_1)\, p(\theta_2|D_2)$$

$$= \int_0^1 \text{Beta}(\theta_1|y_1 + 1, N_1 - y_1 + 1)\, \text{Beta}(\theta_1 - \Delta|y_2 + 1, N_2 - y_2 + 1)\, d\theta_1$$

- Then, evaluate this for any value of Δ

$$P(\Delta > \epsilon|D) = \int_\epsilon^\infty p(\Delta|D)\, d\Delta$$

1. Bayesian decision theory
2. Choosing the “right” model
3. **Frequentist decision theory**
4. Frequentist hypothesis testing

Frequentist decision theory

Frequentist decision theory:

- treat the **unknown state of nature** as a **fixed but unknown quantity**.
- treat the data x as **random**.

Computing the risk of an estimator :

- Risk of an estimator δ given an **unknown state of nature** θ : $R(\theta, \delta) = \mathbb{E}_{p(x|\theta)}[\ell(\theta, \delta(x))]$

Bayes risk :

- **Problem**: the **true state of nature** θ is unknown.
- **Solution**: assume a prior π_0 for θ, **Bayes risk** is given by : $R_{\pi_0}(\delta) = \mathbb{E}_{\pi_0(\theta)}[R(\theta, \delta)] = \int d\theta\, dx\, \pi_0(\theta)\, p(x|\theta)\, l(\theta, \delta(x))$
- A **decision rule** that minimizes the **Bayes risk** is known as a **Bayes estimator**: $\delta(x) = arg\min_a \int d\theta \pi(\theta) p(x|\theta) l(\theta|a)$

Frequentist decision theory

Maximum risk:

- **Problem**: using a prior might be seem **undesirable** in frequentist.
- Define **maximum risk** : $R_{\max}(\delta) = \sup_\theta R(\theta, \delta)$
- **minimax estimator**: minimizes $R_{max}(\delta)$, hard to **compute**.

Consistent estimators

- $\mathbf{x} = \{x_n : n = 1, \ldots, N\}$, $x_n \in \mathcal{X}$ is **iid** sample from $p(X|\theta^*)$
  - $\theta^* \in \Theta$ is true parameter.
- An estimator $\delta : \mathcal{X} \to \Theta$ is **Consistent estimators**.
  - if $\hat{\theta}(x) \to \theta^*$ as $N \to \infty$
- Note that :an estimator can be unbiased but **not consistent.**

Admissible estimators:

- δ_1 dominates δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ.
- An **admissible** estimator is **not strictly dominated** by any other

Empirical risk minimization

In supervised learning:

- **different unknown state of nature** (output y) for each input x.
- Estimator δ is prediction function $\hat{y} = f(x)$, **true** distribution $p^*(x,y)$, **population risk** is given as :

$$R(f,p^*) = R(f) = \mathbb{E}_{p^*(x,y)}(l(y,f(x)))$$

- p^* is unknown, can approximate it as:

$$p_{\mathcal{D}}(x,y|\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(x_n,y_n)\in\mathcal{D}} \delta(x-x_n)\delta(y-y_n)$$

- Plugging this in gives us the **empirical risk**:

$$R(f,D) = \mathbb{E}_{p_D(x,y)}\left[\ell(y,f(x))\right] = \frac{1}{N}\sum_{n=1}^{N}\ell(y_n,f(x_n))$$

Approximation error vs estimation error

Denote:

- $f^{**} = arg\min_f R(f)$, the function achieves the minimal possible **population risk.**
  - we cannot consider **all possible functions.**
- $f^* = arg\min_{f\in\mathcal{H}} R(f)$ is the **best function** in **hypothesis space**,$\mathcal{H}$
  - Cannot compute f^*, since cannot compute **population risk.**
- The prediction function that **minimizes the empirical risk** in **hypothesis space**.

$$f_N = arg\min_{f\in\mathcal{H}} R(f,\mathcal{D}) = arg\min_{f\in\mathcal{H}} \mathbb{E}_{p_{\text{tr}}}[\ell(y,f(x))]$$

We have:

$$\mathbb{E}_{p^*}[R(f_N) - R(f^{**})] = \underbrace{R(f^*) - R(f^{**})}_{\epsilon_{app}(\mathcal{H})} + \underbrace{\mathbb{E}_{p^*}[R(f_N) - R(f^*)]}_{\epsilon_{est}(\mathcal{H},N)}$$

Approximation error vs estimation error(cont.)

We have:

- $\epsilon_{app}(\mathcal{H})$ is **approximation error**.
- $\epsilon_{est}(\mathcal{H}, N)$ is the **estimation error** or **generalization error**. We can approximate it by :

$$\epsilon_{est}(\mathcal{H}, N) \approx \mathbb{E}_{train}[\ell(y, f_N^*(x))] - \mathbb{E}_{test}[\ell(y, f_N^*(x))]$$

This difference is often called the generalization gap.

Statistical learning theory

Bounding the generalization error

- Denote :
  - Data distribution p^*, Dataset $\mathcal{D}$, size N drawn from p^*
  - $R(h, D) = \frac{1}{N}\sum_{i=1}^{N} I(f(x_i) \neq y_i)$ is the **empirical risk.**
  - $R(h) = \mathbb{E}[I(f(x) \neq y)]$ is the **population risk.**
- $$P\left(\max_{h \in H} |R(h) - R(h, D)| > \epsilon\right) \leq 2\,\dim(H) e^{-2N\epsilon^2}$$

VC dimension:

- **Problem**:If the **hypothesis space** $\mathcal{H}$ is infinite, cannot use $dim(H) = |H|$.
- Can use a **quantity** called the **VC dimension** of the hypothesis class.
  - **hard** to compute.

1. Bayesian decision theory
2. Choosing the "right" model
3. Frequentist decision theory
4. Frequentist hypothesis testing

Frequentist hypothesis testing

Likelihood ratio test:

- $p(H_0) = p(H_1) = 0.5$, and that we use 0-1 loss.Accept H_0 iff $\frac{p(D|H_0)}{p(D|H_1)} > 1$

Simple vs compound hypotheses

- we could integrate out these unknown parameters, as in the **Bayesian approach** :

$$\frac{p(H_0 \mid D)}{p(H_1 \mid D)} = \frac{\int_{\theta \in H_0} p(\theta) p_\theta(D)\, d\theta}{\int_{\theta \in H_1} p(\theta) p_\theta(D)\, d\theta} \approx \frac{\max_{\theta \in H_0} p_\theta(D)}{\max_{\theta \in H_1} p_\theta(D)}$$

Type I vs type II errors

- **Type I**(false negative): reject H_0 when H_0 is true.
  - $\alpha(\mu_0) = p(\text{type I error}) = p(X(\tilde{D}) > x^* \mid \tilde{D} \sim H_0)$
- **Type II**(false positive): accept H_0 when H_0 is false.
  - $\beta(\mu_1) = p(\text{type II error}) = p(X(\tilde{D}) < x^* \mid \tilde{D} \sim H_1)$

Type I vs type II errors(cont.)

Type I vs type II errors

- **power** of a test : $1 - \beta(\mu_1)$ is probability reject H_0, given H_1 true.

(a) (b)

Null hypothesis significance testing (NHST) and p-values

- **Goal**:
  - Test if a simple H_0 is "plausible" given some data.
- define the p-value as :

$$p\text{-val} = \Pr(\text{test}(\tilde{D}) \geq \text{test}(D) \mid \tilde{D} \sim H_0)$$

Smaller values correspond to stronger evidence against H_0.

p-values considered harmful