4: Information and Statistical Inference-1 ①

The story so far...

- ∗ Information = Reduction in uncertainty
- ∗ Uncertainty ≈ Randomness ≈ Entropy
- ∗ When an unknown X is revealed, information revealed equals

Uncertainty before − Uncertainty after = $H(X) - 0 = H(X)$

And now for something completely different...

How much information is revealed about X when Y is revealed?

$\boxed{A}$ Statistical Inference: Hypothesis testing and estimation

Let (X, Y) be jointly distributed with joint distribution P_{XY}. Suppose that X is an unknown and Y is the observed r.v. The conditional distribution $P_{Y|X}(\cdot|x)$ is called a channel in information theory. We will denote it by $W: \mathcal{X} \to \mathcal{Y}$ and abbreviate $W_x = P_{Y|X}(\cdot|x)$.

(For discrete $\mathcal{X}, \mathcal{Y}$, $W(y|x)$ denotes the prob. $P(Y=y|X=x)$.)

→ We can now use $(W_x, x \in \mathcal{X})$ to represent an experiment where x is unknown and $Y \sim W_x$ are observed. The goal is to determine x by observing $Y \sim W_x$.

This is the classic Statistical Inference problem.

In our setting, we assume that X is generated from a fixed distribution P_X. Such a formulation is called a Bayesian Formulation.

* Binary Hypothesis Testing : $|\mathcal{X}|=2$

$\mathcal{H}_0 : Y \sim W_0$

$\mathcal{H}_1 : Y \sim W_1$

Upon observing $Y \sim W_X$, we form an estimate $\hat{X} = g(Y)$. Our goal is to minimize the average probability of error

$$\mathbb{P}(X \neq \hat{X}) = p_X(0)\, \mathbb{P}(g(Y)=1|0) + p_X(1)\, \mathbb{P}(g(Y)=0|1)$$

$$= p_X(0) \sum_{y \in A_0^c} W_0(y) + p_X(1) \sum_{y \in A_0} W_1(y),$$

where $A_0 = \{y : g(y) = 0\}$.

The distribution P_X is called the prior.

The distribution $P_{X|Y}(\cdot|y)$ is called the posterior, which can be computed using the Bayes rule: $P_{X|Y}(x|y) = \frac{P_X(x)\, W(y|x)}{P_Y(y)}$

The induced output distribution P_Y is given by

$$P_Y(y) = \sum_x P_X(x)\, W(y|x) =: W \circ P_X(y) \text{ or } W P_X(y).$$

For a uniform prior, i.e., $p_X(0) = p_X(1) = \frac{1}{2}$,

$$\mathbb{P}(\hat{X} \neq X) = \frac{1}{2} W_0(A_0^c) + \frac{1}{2} W_1(A_0)$$

$$= \frac{1}{2} - \frac{1}{2}\left(W_0(A_0) - W_1(A_0) \right)$$

Note that we can choose any A_0. The least probability of error $P_e^*(\text{unif})$ is given by

$$P_e^*(\text{unif}) = \frac{1}{2} - \frac{1}{2} \max_{A} \left(W_0(A) - W_1(A) \right)$$

$$= \frac{1}{2} \cdot \left(1 - d(W_0, W_1) \right)$$

and is attained by

$$A^* = \{ y : \ W_0(y) \geq W_1(y) \}.$$

The function g is called a test and the function $g^*(y) = \mathbb{1}_{\{W_0(y) < W_1(y)\}}$ is called a Bayes optimal test or simply Bayes.

* M-ary hypothesis testing: $|\mathcal{X}| = M$

Test $g : \mathcal{Y} \to \mathcal{X}$ outputs $\hat{X} = g(Y)$.

$P_e(g | P_X) = \mathbb{P}(\hat{X} \neq X)$

Optimal prob. of error $\equiv P_e^*(P_X) = \min_{g : \mathcal{Y} \to \mathcal{X}} P_e(g | P_X)$.

* Estimation problem: $\mathcal{X}$ need not be discrete.

Note that $\mathbb{P}(\hat{X} \neq X) = \mathbb{E}_{P_{XY}} \left[\mathbb{1}_{\{\hat{X} \neq X\}} \right]$.

For a general $\mathcal{X}$, we can consider an arbitrary loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Then, $L(g | P_X) := \mathbb{E}_{P_{XY}} \left[\ell(g(Y), X) \right]$.

and optimal loss $L^*(P_X) = \min_{g:\mathcal{Y}\to\mathcal{X}} L(g|P_X)$.

When $\mathcal{X} = \mathbb{R}^d$, $\ell(x,x') = \|x-x'\|_2^2 = \sum_{i=1}^{d} (x_i - x_i')^2$ is a popular loss function, and $L(g/P_X)$ is called the Mean Squared Loss (MSE). The quantity $L^*(P_X)$ in this case is called the Minimum Mean Squared Error (MMSE).

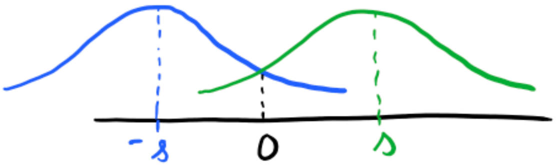
B Examples

* Example 1 (Binary hypothesis testing)

$\mathcal{X} = \{-s, s\}$, $\mathcal{Y} = \mathbb{R}$, $\mathcal{H}_0 : Y \sim N(-s, \sigma^2)$, $\mathcal{H}_1 : Y \sim N(s, \sigma^2)$

Then, $P_e^*(\text{unif}) = \frac{1}{2}\left(1 - d(W_{-s}, W_s)\right)$

$$\text{where } d(W_{-s}, W_s) = W_{-s}(\{y : W_{-s}(y) > W_s(y)\}) - W_s(\{y : W_{-s}(y) \le W_s(y)\})$$



$$W_{-s}(A^*) = \mathbb{P}(-s + Z < 0) \quad \text{where } Z \sim N(0, \sigma^2)$$
$$= 1 - \mathbb{P}(Z \geq s)$$
$$= 1 - Q\left(\frac{s}{\sigma}\right) \quad \text{where } Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-t^2/2}\, dt.$$

$$W_s(A^*) = \mathbb{P}(s + Z < 0) = \mathbb{P}(Z < -s) = Q\left(\frac{s}{\sigma}\right)$$ (why?)

Therefore, $d(W_{-s}, W_s) = 1 - 2Q\left(\frac{s}{\sigma}\right)$, and so

$$P_e^*(\text{unif}) = Q\left(\frac{s}{\sigma}\right) \approx e^{-\frac{s^2}{2\sigma^2}}$$

* Example 2 $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $W_x = N(x, \sigma^2 I)$

↳ Gaussian with mean x, covariance $\sigma^2 I$

Then, for $g(y) = y$, we have

$L(g|P_x) = d\sigma^2$ for any P_x (show this)

for the MSE $L(g|P_x)$.

* Example 3 (Testing the bias of a coin)

How many coin tosses are needed to test if a coin is head heavy or tail heavy?

In our framework, $\mathcal{X} = \{\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\}$ and $\mathcal{Y} = \{0,1\}^n$.

$\mathcal{H}_0: Y_1, \dots, Y_n \sim \text{iid Ber}\left(\frac{1}{2} - \varepsilon\right) = P$

$\mathcal{H}_1: Y_1, \dots, Y_n \sim \text{iid Ber}\left(\frac{1}{2} + \varepsilon\right) = Q$

Let $P^n = P \times \dots \times P$ and $Q^n = Q \times \dots \times Q$ denote the n-fold product distributions on $Y_1, \dots, Y_n$.

Then, by our formula for $P_e^*(\text{unif})$,

$P_e^*(\text{unif}) = \frac{1}{2}\left(1 - d(P^n, Q^n)\right)$

Suppose we will be happy with $P_e^*(\text{unif}) \le \frac{1}{2000}$. Then, we must have $d(P^n, Q^n) \geq \frac{999}{1000}$.

Heuristically, $d(P^n, Q^n)$ grows with n towards 1. The least number of coin flips we need is the least n required for $d(P^n, Q^n)$ to cross $999/1000$.

To find an estimate for this least n, we will derive an upper bound for $d(P^n, Q^n)$.

Lemma (Subadditivity of total variation distance)

For $P^n = P_1 \times P_2 \times \ldots \times P_n$ and $Q^n = Q_1 \times Q_2 \times \ldots \times Q_n$, we have

$$d(P^n, Q^n) \leq \sum_{i=1}^{n} d(P_i, Q_i).$$

Proof. It suffices to show the claim for $n = 2$ (why?).

$$\begin{aligned}
d(P_1 \times P_2, Q_1 \times Q_2) &= \frac{1}{2} \sum_{y_1, y_2} \left| P_1(y_1) P_2(y_2) - Q_1(y_1) Q_2(y_2) \right| \\
&= \frac{1}{2} \sum_{y_1, y_2} \Big| P_1(y_1) P_2(y_2) - P_1(y_1) Q_2(y_2) \\
&\qquad + P_1(y_1) Q_2(y_2) - Q_1(y_1) Q_2(y_2) \Big| \\
&\leq \underbrace{\frac{1}{2} \sum_{y_1, y_2} \left| P_1(y_1) P_2(y_2) - P_1(y_1) Q_2(y_2) \right|}_{\frac{1}{2} \sum_{y_1} P_1(y_1) \sum_{y_2} \left| P_2(y_2) - Q_2(y_2) \right| = d(P_2, Q_2)} \\
&\qquad + \underbrace{\frac{1}{2} \sum_{y_1, y_2} \left| P_1(y_1) Q_2(y_2) - Q_1(y_1) Q_2(y_2) \right|}_{d(P_1, Q_1)} \\
&= d(P_2, Q_2) + d(P_1, Q_1). \qquad \blacksquare
\end{aligned}$$

Thus, for $P_e^*(\text{unif}) \le \frac{1}{2000}$, we need $n \ge \frac{999}{1000} \cdot \frac{1}{d(P,Q)}$.

For the coin-toss example, $d(P,Q) = \varepsilon$. Then, the previous bound suggests that we need $\gtrsim 1/\varepsilon$ coin tosses to distinguish the two coins. In fact, we will see later that this bound is weak and we need many more coin tosses, roughly $1/\varepsilon^2$. We will see that $d(P^n, Q^n) \lesssim \sqrt{n}$ and not just $\lesssim n$ as suggested by the lemma above.

C Neyman-Pearson formulation and threshold tests

The Bayes optimal test we saw had the form

$$g(y) = \begin{cases} 0, & \frac{W_0(y)}{W_1(y)} > 1 \\ 1, & \frac{W_0(y)}{W_1(y)} \le 1 \end{cases}$$

$$= \begin{cases} 0, & \log \frac{W_0(y)}{W_1(y)} > 0 \\ 1, & \log \frac{W_0(y)}{W_1(y)} \le 0 \end{cases}$$

This suggests the following class of tests:

$$g_\tau(y) = \begin{cases} 0, & \log \frac{W_0(y)}{W_1(y)} > \tau \\ 1, & \log \frac{W_0(y)}{W_1(y)} \le \tau \end{cases}$$

How well do these tests perform?

$$P_e(g_\tau | P_X) = P_{X(0)} \sum_{y:\, g_\tau(y)=1} W_0(y) \; + \; P_X(1) \sum_{y:\, g_\tau(y)=0} W_1(y)$$

Abbreviating $P_{X(1)} = p$,

$$P_e(g_\tau | P) = (1-p)\, \underbrace{W_0\left(\{g_\tau(Y)=1\}\right)}_{\substack{\text{Error given } X=0 \\ \text{type I error}}} + \; p\, \underbrace{W_1\left(\{g_\tau(Y)=0\}\right)}_{\substack{\text{Error given } X=1 \\ \text{type II error}}}$$

As τ increases, type-I error increases and type-II error decreases.

Neyman-Pearson considered a slighty different formulation than the average error criterion above.

We seek tests for which error of type-I is less than ε. Under this constraint, we want to find a test that minimizes the error of type-II. Namely, find a test that attains

$$\beta_\varepsilon(W_0, W_1) = \min \Big\{ \sum_{y:\, g(y)=0} W_1(y) : \text{ test } g \text{ satisfies } \sum_{y:\, g(y)=1} W_0(y) \le \varepsilon \Big\}$$

→ This is motivated by applications where $\mathcal{H}_0$ is the normal operation and $\mathcal{H}_1$ is an alarming situation. Here, the error of type-I is a false alarm and is a less severe error, while the error of type-II is a missed detection and is a more severe error. The Neyman-Pearson formulation seeks to minimize the missed detection probability given that the prob. of false alarm is less than ε.

→ We evaluate our threshold test g_τ for this setting.

Prob. of false alarm

$$= \text{Prob. of error of type-I} = W_0\left(\{y : g_\tau(y)=1\}\right)$$

$$= W_0\left(\left\{y : \log\frac{W_0(y)}{W_1(y)} \le \tau\right\}\right)$$

Prob. of missed detection $= W_1\left(\{y : g_\tau(y)=0\}\right)$

$$= \sum_{y:\ \log\frac{W_0(y)}{W_1(y)} > \tau} W_1(y) \rightsquigarrow = \frac{W_1(y)}{W_0(y)} \cdot W_0(y)$$

$$= \left(2^{-\log\frac{W_0(y)}{W_1(y)}}\right) \cdot W_0(y)$$

$$< \sum_{y} 2^{-\tau} W_0(y)$$

$$= 2^{-\tau}.$$

We have shown the following result:

Lemma Suppose that $\lambda > 0$ satisfies

$$W_0\left(\left\{y : \log\frac{W_0(y)}{W_1(y)} \geq \tau\right\}\right) \geq 1-\varepsilon.$$

Then,

$$\beta_\varepsilon(W_0, W_1) \le 2^{-\tau}.$$

For iid observation $y = y^n$

$$W_0(y^n) = \prod_{i=1}^{n} p(y_i), \quad W_1(y^n) = \prod_{i=1}^{n} q(y_i)$$

Then, by Chebyshev's inequality,

$$P\left(\sum_{i=1}^{n} \log \frac{p(Y_i)}{q(Y_i)} \le n\mathbb{E}_P\left[\log \frac{p(Y_1)}{q(Y_1)}\right] + \sqrt{\frac{n}{\varepsilon} \mathrm{Var}\left(\log \frac{p(Y_1)}{q(Y_1)}\right)}\right)$$
$$\geq 1 - \varepsilon.$$

$\boxed{D}$ KL Divergence and Stein's lemma

The quantity $\mathbb{E}_P\left[\log \frac{p(Y)}{q(Y)}\right]$ is called the Kullback-Leibler Divergence and is denoted by $D(P\|Q)$.

$D(P\|Q) = \sum_y p(y) \log \frac{p(y)}{q(y)}$ is the counterpart of $d(P,Q)$ that enters the Neyman-Pearson formulation.

Our lemma in part C shows:

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_\varepsilon(P^n, Q^n) \geq D(P\|Q)$$

Stein's lemma

$$\boxed{\lim_{n\to\infty} -\frac{1}{n} \log \beta_\varepsilon(P^n, Q^n) = D(P\|Q),}$$

Namely, the largest exponential decay rate of $\beta_\varepsilon(P^n, Q^n)$ is $D(P\|Q)$ and is attained by threshold tests.

$D(P\|Q)$ has a similar interpretation as $d(P,Q)$: if $D(P\|Q)$ is small, the hypotheses P and Q are difficult to distinguish. S 'n's lemma gives an asymptotic justification of this fact. Lat we will see another justification that holds for a fixed n.

Example $P \equiv \mathrm{Ber}\left(\frac{1}{2}\right)$, $Q \equiv \mathrm{Ber}\left(\frac{1+\varepsilon}{2}\right)$

$$d(P,Q) = \frac{\varepsilon}{2}$$

$$D(P\|Q) = \frac{1}{2}\log\frac{1}{1+\varepsilon} + \frac{1}{2}\log\frac{1}{1-\varepsilon} = \frac{1}{2}\log\frac{1}{1-\varepsilon^2}$$

$$= \frac{1}{2\ln 2}\ln\frac{1}{1-\varepsilon^2} \geq \frac{\varepsilon^2}{2\ln 2} = \frac{2}{\ln 2}d(P,Q)^2$$

(why?)

* Continuous distributions

For P and Q with densities f and g,

$$D(P\|Q) := \int f(x)\log\frac{f(x)}{g(x)}\,dx.$$

This definition serves exactly the same purpose as that for the discrete case (In fact, both can be recovered as special cases of a more general definition)

The log-likelihood ratio tests g_τ can now be replaced with $\sum_{i=1}^{n}\log f(x_i)/g(x_i) \overset{0}{\underset{1}{\gtrless}} \tau$, with the same performance.

E Properties of KL divergence (proofs will be given later)

(1) Data processing inequality

"Distances b/w distributions decreases when you further process their samples"

- Agrees with our heuristic that these distances determine how difficult is it to test between the two distributions. (since we can apply tests to the processed samples)

Let P and Q be two distributions on $\mathcal{Y}$, and let $W: \mathcal{Y} \to \mathcal{Z}$ be a fixed channel (representing the data processing operation).

Then,

$$\text{(i)} \quad d(W\circ P, W\circ Q) \leq d(P, Q)$$

$$\text{(ii)} \quad D(W\circ P, W\circ Q) \leq D(P\|Q)$$

(2) Pinsker's inequality

(The bound of our example is tight)

$$D(P\|Q) \geq \frac{2}{\ln 2}\, d(P,Q)^2$$

This bound says that $D(P\|Q)$ behaves roughly the same as the square of distance (what is special about squared distance in Euclidean space?)

(3) Additivity $\quad D(P_1 \times \dots \times P_n \| Q_1 \times \dots \times Q_n) = \sum_{i=1}^{n} D(P_i \| Q_i)$