パターン認識
Pattern Recognition

佐藤真一
Shin'ichi Satoh

国立情報学研究所
National Institute of Informatics

May 16, 2023

On Assignments

- Since there was system trouble on ITC LMS on May 9 and 10, the deadline of the assignment issued on April 25 has been extended from May 9 to May 16, by one week.
- Thus forthcoming deadlines are as follows.
  - Assignment issued on April 25, due on May 9 now extended to May 16
  - Assignment issued on May 2, due on May 16
  - No assignment on May 9
  - Assignment issued on May 16 (today), due on May 30
- PDF format file should be submitted to ITC LMS
- IMPORTANT: Don't submit fines in Jupyter Notebook (ipynb) or Python (py)!

Today's Roadmap

- Error rate estimation from samples
- The Resubstitution (R) method
- The Holdout (H) method
- The Cross-Validation (CV) method
- The bootstrap method

Estimation of Classificaiton Errors

For simplicity, we think about two-class classification problem. A classifier is expressed by

$$h(X) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$$

where $h(X)$ is the discriminant function of a vector X. The *probability of error* for this classifier is

$$\varepsilon_1 = \int_{h(X)>0} p_1(X)dX = \int u(h(X))p_1(X)dX$$

where $u(\cdot)$ is the step function and $p_i(X)$ is the distribution of the class i.

Let's consider Fourier and inverse Fourier transform

$$\mathscr{F}[x(t)] = X(\omega) = \int_{-\infty}^{\infty} x(t)\mathrm{e}^{-j\omega t}dt$$

$$\mathscr{F}^{-1}[X(\omega)] = x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\mathrm{e}^{j\omega t}d\omega$$

Step function $u(t)$ is defined as follows:

$$u(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2} & t = 0 \\ 1 & t > 0 \end{cases}$$

Its Fourier transform

$$\begin{aligned}\mathscr{F}[u(t)] &= \frac{1}{2}\mathscr{F}[1] + \int_0^\infty \mathrm{e}^{-j\omega t} dt \\ &= \frac{1}{2}(2\pi\delta(\omega)) - \frac{1}{j\omega}\left[\mathrm{e}^{-j\omega t}\right]_0^\infty \\ &= \pi\delta(\omega) + \frac{1}{j\omega}\end{aligned}$$

and its inverse Fourier transform

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} [\pi\delta(\omega) + \frac{1}{j\omega}]\mathrm{e}^{j\omega t} d\omega$$

Therefore

$$\begin{aligned}\varepsilon_1 &= \int_{h(X)>0} p_1(X)dX = \int u(h(X))p_1(X)dX \\ &= \frac{1}{2\pi}\iint [\pi\delta(\omega) + \frac{1}{j\omega}]\mathrm{e}^{j\omega h(X)}p_1(X)d\omega dX \\ &= \frac{1}{2} + \frac{1}{2\pi}\iint \frac{\mathrm{e}^{j\omega h(X)}}{j\omega}p_1(X)d\omega dX\end{aligned}$$

Likewise

$$\varepsilon_2 = \int_{h(X)<0} p_2(X)dX = \frac{1}{2} - \frac{1}{2\pi}\iint \frac{e^{j\omega h(X)}}{j\omega} p_2(X)d\omega dX.$$

Then the total probability of error is

$$\begin{aligned}\varepsilon =& P_1\varepsilon_1 + P_2\varepsilon_2 \\ =& \frac{1}{2} + \frac{1}{2\pi}\iint \frac{e^{j\omega h(X)}}{j\omega}\tilde{p}(X)d\omega dX\end{aligned}$$

where

$$\tilde{p}(X) = P_1 p_1(X) - P_2 p_2(X).$$

Error Estimation by Samples

We show the justification of *error-counting procedure* when only finite number of samples available.
$p_i(X)$ may be replaced by

$$\hat{p}_i(X) = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta(X - \mathbf{X}_j^{(i)})$$

where $\mathbf{X}_j^{(i)}$ are N_i test samples drawn from $p_i(X)$.

Then the estimate of the error probability is

$$\begin{aligned}\hat{\varepsilon} =& \frac{1}{2} + \frac{1}{2\pi} \iint \frac{e^{j\omega h(X)}}{j\omega} [\frac{P_1}{N_1} \sum_{j=1}^{N_1} \delta(X - \mathbf{X}_j^{(1)}) - \frac{P_2}{N_2} \sum_{j=1}^{N_2} \delta(X - \mathbf{X}_j^{(2)})] d\omega dX \\ =& \frac{1}{2} + \frac{P_1}{N_1} \sum_{j=1}^{N_1} \boldsymbol{\alpha}_j^{(1)} - \frac{P_2}{N_2} \sum_{j=1}^{N_2} \boldsymbol{\alpha}_j^{(2)}\end{aligned}$$

where

$$\boldsymbol{\alpha}_j^{(i)} = \frac{1}{2\pi} \int \frac{e^{j\omega h(\mathbf{X}_j^{(i)})}}{j\omega} d\omega = \frac{sign(h(\mathbf{X}_j^{(i)}))}{2}$$

Then

$$\frac{1}{N_1}\sum_{j=1}^{N_1}\boldsymbol{\alpha}_j^{(1)} = \frac{1}{2N_1}[(\#\ \omega_1\text{-errors}) - (\#\ \omega_1\text{-corrects})]$$
$$= \frac{1}{N_1}(\#\ \omega_1\text{-errors}) - \frac{1}{2}.$$

Likewise,

$$\frac{1}{N_2}\sum_{j=1}^{N_2}\boldsymbol{\alpha}_j^{(2)} = -\frac{1}{N_2}(\#\ \omega_2\text{-errors}) + \frac{1}{2}.$$

Putting all together,

$$\hat{\varepsilon} = P_1\frac{(\#\ \omega_1\text{-errors})}{N_1} + P_2\frac{(\#\ \omega_2\text{-errors})}{N_2}.$$

Mean and Variance of the Estimated Error

Let's derive the mean and variance (w.r.t. test sample) of the estimated error:

$$
\begin{aligned}
E_t\{\boldsymbol{\alpha}_j^{(i)}\} =& \bar{\alpha}_i = \frac{1}{2\pi} \iint \frac{\mathrm{e}^{j\omega h(X)}}{j\omega} p_i(X) d\omega dX \\
=& \begin{cases} \varepsilon_1 - \dfrac{1}{2} & (i = 1) \\ \dfrac{1}{2} - \varepsilon_2 & (i = 2) \end{cases} \\
E_t\{\boldsymbol{\alpha}_i^{(i)^2}\} =& E_t\{[\frac{1}{2\pi} \int \frac{\mathrm{e}^{j\omega h(\mathbf{X})}}{j\omega} d\omega]^2\} = E_t\{[\frac{1}{2} sign(h(\mathbf{X}))]^2\} \\
=& \frac{1}{4} \\
E_t\{\boldsymbol{\alpha}_j^{(i)} \boldsymbol{\alpha}_\ell^{(k)}\} =& \bar{\alpha}_i \bar{\alpha}_k \quad (i \neq k \text{ or } j \neq \ell)
\end{aligned}
$$

Mean and Variance of the Estimated Error

Putting these together:

$$
\begin{aligned}
E_t\{\hat{\varepsilon}\} =& \frac{1}{2} + P_1\bar{\alpha_1} - P_2\bar{\alpha_2} \\
=& \frac{1}{2} + P_1(\varepsilon_1 - \frac{1}{2}) - P_2(\frac{1}{2} - \varepsilon_2) = \varepsilon \\
Var_t\{\hat{\varepsilon}\} =& \frac{P_1^2}{N_1} Var_t\{\boldsymbol{\alpha}_j^{(1)}\} + \frac{P_2^2}{N_2} Var_t\{\boldsymbol{\alpha}_j^{(2)}\} \\
=& \frac{P_1^2}{N_1}[\frac{1}{4} - (\varepsilon_1 - \frac{1}{2})^2] + \frac{P_2^2}{N_2}[\frac{1}{4} - (\frac{1}{2} - \varepsilon_2)^2] \\
=& P_1^2 \frac{\varepsilon_1(1-\varepsilon_1)}{N_1} + P_2^2 \frac{\varepsilon_2(1-\varepsilon_2)}{N_2}.
\end{aligned}
$$

Finally, $\hat{\varepsilon}$ is an unbiased and consistent estimate irrespective of $h(X)$.

Another solution

Let $\hat{\tau}_i$ be the number of misclassified samples in class i. Then the random variables $\hat{\tau}_1$ and $\hat{\tau}_2$ are independent and yielding binomial distribution:

$$\begin{aligned} Pr\{\hat{\tau}_1 = \tau_1, \hat{\tau}_2 = \tau_2\} &= \prod_{i=1}^{2} Pr\{\hat{\tau}_i = \tau_i\} \\ &= \prod_{i=1}^{2} \left(\begin{array}{c} N_i \\ \tau_i \end{array} \right) \varepsilon_i^{\tau_i} (1 - \varepsilon_i)^{N_i - \tau_i}. \end{aligned}$$

The ω_i-error, ε_i, can be estimated by $\dfrac{\hat{\tau}_i}{N_i}$:

$$\hat{\varepsilon} = \sum_{i=1}^{2} P_i \frac{\hat{\tau}_i}{N_i}.$$

Another solution

By using the mean and variance of binomial distribution,

$$E\{\hat{\varepsilon}\} = P_1\varepsilon_1 + P_2\varepsilon_2 = \varepsilon$$

$$Var\{\hat{\varepsilon}\} = P_1^2\frac{\varepsilon_1(1-\varepsilon_1)}{N_1} + P_2^2\frac{\varepsilon_2(1-\varepsilon_2)}{N_2}$$

Bounds of the Bayes Error

When only a finite number of samples are available for training and testing, we need to take a compromise: which samples are for training and which are for testing.
We represent the classification error as a function of two sets, the design (training) and test sets:

$$\varepsilon(\mathscr{P}_D, \mathscr{P}_T)$$

where $\mathscr{P}$ is a set of densities of classes, e.g.,

$$\mathscr{P} = \{p_1(X), p_2(X)\}.$$

Bounds of the Bayes Error

If the slassifier is the Bayes for the given test distributions, the resulting error is minimum:

$$\varepsilon(\mathscr{P}_T, \mathscr{P}_T) \leq \varepsilon(\mathscr{P}_D, \mathscr{P}_T).$$

The Bayes error for the true $\mathscr{P}$ is $\varepsilon(\mathscr{P}, \mathscr{P})$. Since we never know the true $\mathscr{P}$, let's try to find the upper and lower bounds of $\varepsilon(\mathscr{P}, \mathscr{P})$ by using its estimate obtained by finite number of samples $\hat{\mathscr{P}} = (\hat{\mathbf{p}}_1(X), \hat{\mathbf{p}}_2(X))$:

$$\begin{aligned}\varepsilon(\mathscr{P}, \mathscr{P}) &\leq \varepsilon(\hat{\mathscr{P}}, \mathscr{P}) \\ \varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}) &\leq \varepsilon(\mathscr{P}, \hat{\mathscr{P}}).\end{aligned}$$

Since we know that the error counting procedure is unbiased estimate,

$$\varepsilon(\hat{\mathscr{P}}, \mathscr{P}) = E_{\hat{\mathscr{P}}_T}\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)\}$$

where $\hat{\mathscr{P}}_T$ is another set independent of $\hat{\mathscr{P}}$.

Recap:

$$\begin{aligned}\varepsilon(\mathscr{P}, \mathscr{P}) \leq & \varepsilon(\hat{\mathscr{P}}, \mathscr{P}) \\ \varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}) \leq & \varepsilon(\mathscr{P}, \hat{\mathscr{P}}) \\ \varepsilon(\hat{\mathscr{P}}, \mathscr{P}) = & E_{\hat{\mathscr{P}}_T}\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)\}.\end{aligned}$$

Similarly,

$$E\{\varepsilon(\mathscr{P}, \hat{\mathscr{P}})\} = \varepsilon(\mathscr{P}, \mathscr{P}).$$

Putting these together,

$$E\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}})\} \leq \varepsilon(\mathscr{P}, \mathscr{P}) \leq E_{\hat{\mathscr{P}}_T}\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)\}.$$

Holdout (H) Method

$\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)$ can be obtained by two independent sample sets, $\hat{\mathscr{P}}$ and $\hat{\mathscr{P}}_T$ from $\mathscr{P}$, and using $\hat{\mathscr{P}}$ for designing the Bayes classifier and $\hat{\mathscr{P}}_T$ for testing.
As we observed, $E_{\hat{\mathscr{P}}_T}\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)\}$ gives the upper bound of the Bayes error.
Similarly, $E_{\hat{\mathscr{P}}}\{E_{\hat{\mathscr{P}}_T}\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}}_T)\}\}$ gives the upper bound.
This procedure is called the holdout (H) method.

The Holdout (H) Method

1. Given data set S is decomposed into two disjoint sets, the design set S_D and the test set S_T.
2. Train classifier using S_D.
3. Estimate the performance using S_T.

Resubstitution (R) Method

$E\{\varepsilon(\hat{\mathscr{P}}, \hat{\mathscr{P}})\}$ gives the lower bound of the Bayes error.
This can be obtained by using $\hat{\mathscr{P}}$ for designing the Bayes classifier and the same $\hat{\mathscr{P}}$ for testing.
This procedure is called the resubstitution (R) method.

The Resubstitution (R) Method

1. Given data set S is used for training classifier.
2. The same set is then used for estimating the performance.

Cross-Validation (CV) Method

Since the holdout method needs to separate the data into design and test set, it has drawback in estimating the error rate.
The bias of the estimation is effected by the size of design set, and the variance is effected by the size of test set.
To alleviate this drawback, the cross-validation (CV) method is also used.

The Cross-Validation (CV) Method

1. Given data set S is separated into N disjoint sets, $S_1, S_2, \cdots, S_N$.
2. Select one set as the test set and combine the remaining $N-1$ sets as the design set, and then perform training and testing.
3. The above procedure is repeated N times by changing the test set, and the resulting N error rates are then averaged.

Cross-Validation (CV) Method

In CV method, you can treat each item as an independent set.
This extreme case is called the leave-one-out (L) method.

Bootstrap Method

In estimating the Bayes error rate ε, we can obtain the estimate $\hat{\varepsilon}$ using R method.
We then can think of the bias:

$$b = \varepsilon - \hat{\varepsilon}.$$

If we can well estimate b, we can estimate the Bayes error by:

$$\varepsilon = \hat{\varepsilon} + b.$$

Bootstrap Method

The bootstrap method generate a pseudo sample set S^* from the original set S by using sampling with replacement (note that $|S| = |S^*|$).
We then estimate the bias b by

$$b^* = \varepsilon^* - \hat{\varepsilon}^*$$

where ε^* is the estimated error by using S^* for training and S for testing, and $\hat{\varepsilon}^*$ by using S^* for training and testing.
We repeat the above procedure B times using B pseudo sample sets and take average to obtain b^*.
The estimated Bayes error is then obtained by

$$\varepsilon = \hat{\varepsilon} + b^*.$$

Bootstrap Method

The Bootstrap Method

1. Obtain the estimated error $\hat{\varepsilon}$ by using the original set S for training and testing (R method).
2. Generate a pseudo sample set S^* by $|S|$-times sampling with replacement from S.
3. Obtain the estimated error rate ε^* by using S^* for training and S for testing.
4. Obtain the error rate $\hat{\varepsilon}^*$ by using S^* for training and testing.
5. Obtain the estimated bias $b^* = \varepsilon^* - \hat{\varepsilon}^*$.
6. The above procedure is repeated B times by changing the pseudo set and the resulting B biases are then averaged to obtain final b^*.
7. The estimated error is obtained by $\varepsilon = \hat{\varepsilon} + b^*$.

Exercise

Two normal distributions

Class 1 $\Sigma_1 = I$, $\mu_1 = [s, 0, \cdots, 0]^T$

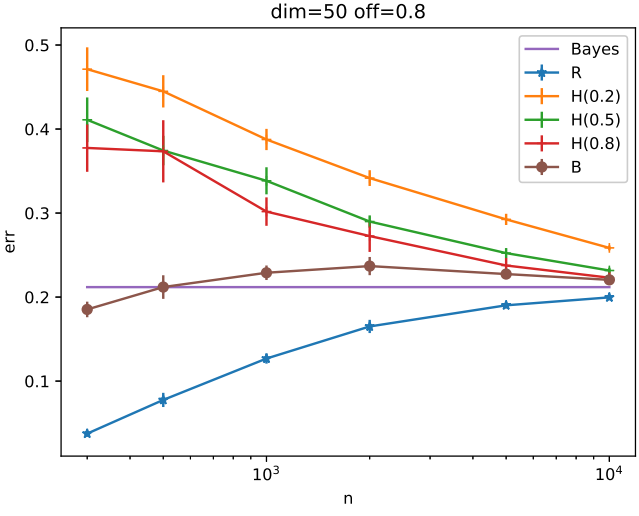
Class 2 $\Sigma_2 = I$, $\mu_2 = [-s, 0, \cdots, 0]^T$

Bayes error:

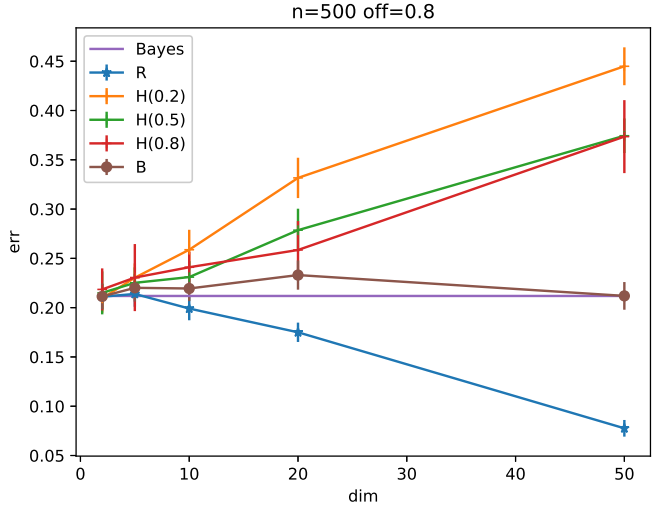
$$P^* = \int_s^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty N(0, I) dx_1 \cdots dx_n$$
$$= \frac{1}{2}(1 - \mathrm{erf}(\frac{s}{\sqrt{2}}))$$
$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Given n class 1 and n class 2 data, observe the error rate of quadratic classifier by using the resubstitution (R) and the holdout (H), and compare them with the Bayes error.

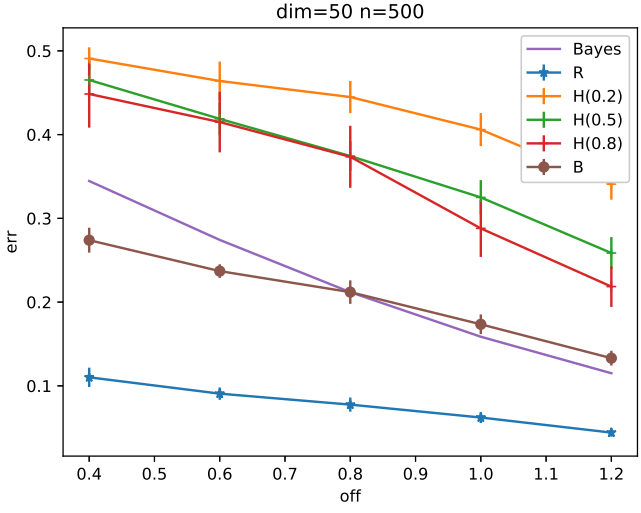
Exercise



Exercise


n=500 off=0.8
Bayes
R
H(0.2)
H(0.5)
H(0.8)
B
err
dim
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0
10
20
30
40
50

Exercise


dim=50 n=500
Bayes
R
H(0.2)
H(0.5)
H(0.8)
B
err
0.5
0.4
0.3
0.2
0.1
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
off

Assignment (Programming project)

- Implement the bootstrap method.
- Following the exercise, compare the estimated error by using the resubstitution (R), the holdout (H), and the bootstrap methods, along with the Bayes error rate.
- Plot the estimation by changing the dimension, the Bayes error rate, and the number of samples.
- Due on May 30.
- PDF format file should be submitted to ITC LMS
- IMPORTANT: Don't submit fines in Jupyter Notebook (ipynb) or Python (py)!