## Week 9.3: Evaluation in ML

**Objective**:1) Is model A better than B ? 2) Is the difference between the results statistically significant?

**A/B Testing**: Simple Control Experiments: 1) Randomly split traffic between two or more version eg.(A) Control, (B) Treatment (statistical test) เพื่อ confirm ความต่างเมื่อมี feature ใหม่

$$var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

Hypothesis Testing: สำรวจ 2 สมมติฐาน H0, HA เพื่อหาความแตกต่าง

Type I - reject H0 when in fact it True.(FP) "False Alarm"

Type II - FTR H0 when in fact False.(FN)

P-value - ถ้า p-value <= alpha: then reject H0 at level alpha

ทำ p-value testing 1.calculate test statistic

2. Convert the result to a p-value by comparing its value to the distribution of test statistics under the null hypothesis.

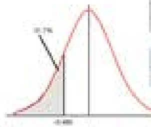
3.Decide ว่า reject หรือไม่ reject H0.

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{var(A)}{n_A} + \frac{var(B)}{n_B}}} \qquad t = \frac{\bar{X}_D \times \sqrt{n}}{\sigma_D}$$

ดูตารางตามค่า significant ที่ได้ ถ้าเป็นเล็กๆ ก็เอาไป -1 ด้วย

unpair

| | Team A | Team B |
|---|---|---|
| N | 10 | 15 |
| Mean | 20.600 | 23.800 |
| Std Dev | 14.017 | 15.485 |
| Variance | 196.489 | 239.844 |



| | |
|---|---|
| Test statistic | t = -0.4850 |
| P-Value | P(T≤t) one tail = 0.317 |

one tail

มากกว่า alpha ยอมรับ H0

paired(มี D(delta))

$$t = \frac{\bar{X}_D \times \sqrt{n}}{\sigma_D}$$

Look at the mean and standard deviation of the differences (delta)

$$t = \frac{-3.2 \times \sqrt{10}}{5.2} = -1.946$$

| | |
|---|---|
| Observations (n) | 10 |
| Mean of differences (delta) | -3.2 |
| Std Dev of differences (delta) | 5.20 |

| | |
|---|---|
| Test statistic | t = -1.946 |
| P-Value | P(T≤t) = 0.084 |

two tails

### Difference in proportions

A t-test is sometimes used to analyse differences in proportions e.g. comparison of conversion rates in A/B testing.

Requires a number of assumptions about the population which are usually not true.

| | Control | Treatment |
|---|---|---|
| Samples | $n_1$ | $n_2$ |
| Conversions | $x_1$ | $x_2$ |

$$p = \frac{x1 + x2}{n1 + n2} \quad p_1 = \frac{x_1}{n_1} \quad p_2 = \frac{x_2}{n_2}$$

$$t\ \text{statistic} = \frac{\text{Difference in proportions}}{\text{Standard error}} \qquad t = \frac{p1 - p2}{\sqrt{p(1-p) \times (\frac{1}{n_1} + \frac{1}{n_2})}}$$

### McNemar's Test

Measure for comparing paired proportions.

e.g. Which is better, classifier C2 or C3 ?

Applied to 2x2 contingency table.

Test captures two key differences:

$n_{01}$: number misclassified by 1st but not 2nd classifier.

$n_{10}$: number misclassified by 2nd but not 1st classifier.

Contingency for C2 v C1

| | |
|---|---|
| 3 $n_{00}$ | 2 $n_{01}$ |
| 0 $n_{10}$ | 5 $n_{11}$ |

McNemar C2 v C1 $\chi^2 = 1/2 = 0.5$

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Note: For test to be applicable require $(n_{01}+n_{10}) > 10$

Contingency for C3 v C1

| | |
|---|---|
| 1 $n_{00}$ | 2 $n_{01}$ |
| 4 $n_{10}$ | 3 $n_{11}$ |

McNemar C3 v C1 $\chi^2 = 1/6 = 0.1666$

- $\chi^2 > 3.84$ required for statistical significance at 95%. So neither classifier significantly better!

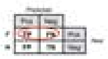
## Week 10.1: Decision Analytic Thinking

Acc = 1- error rate หรือ Correct/Total Predict

Misclassification rate = 1 - Acc

True Positive Rate:
Focus on TPs
Also called Sensitivity
or called Recall, aka. Probability of Detection

$$TPRate = \frac{TP}{TP + FN}$$



False Positive Rate:
Focus on FPs

$$FPRate = \frac{FP}{FP + TN}$$

True Negative Rate:
Focus on TNs
Also called Specificity

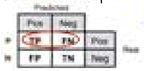
$$TNRate = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

ค่า sensitive คือ จะไม่(FP, FN) ถ้ามีค่าเยอะ ค่า recall หรือ precision

มีtrade off ระหว่าง precision & Rec

$$Recall = \frac{TP}{TP + FN} = Sensitivity$$



$$\text{Balanced Accuracy (Rate)} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}\right)$$

balance ระหว่าง recall กับ TN rate

**Balanced Error Rate (BER)**

$$\text{Balanced Error Rate} = \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{TP+FN}\right)$$

**F-Measure:** A single measure that trade off precision against recall, for a given level of balance.(เขียน precision กับ recall ให้อยู่ในสูตรเดียว) โดยมีเบต้าเป็นพารามิเตอร์(beta = 1: เป็น f1): beta < 1: focus more on precision, >1: recall, =1:harmonic mean on prec and rec.

$$F = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

$$EV = p(o_1) \times v(o_1) + p(o_2) \times v(o_2) + p(o_3) \times v(o_3) + ..$$

- **Online marketing:**
- Expected benefit of targeting $= p_R(x) \times v_R + [1 - p_R(x)] \times v_{NR}$
- Product Price: \$200
- Product Cost: \$100
- Targeting Cost: \$1

$$p_R(x) \times \$99 - [1 - p_R(x)] \times \$1 > 0$$
$$p_R(x) > 0.01$$

PR(X)*99 + (-1-PR(X))*(-1)

$$p(x, y) = p(y) \times p(x \mid y)$$

$T = 110$

$P = 61$ $\quad N = 49$

$p(p) = 0.55$ $\quad p(n) = 0.45$

$p(Y|p) = 56/61 = 0.92$ $\quad p(Y|n) = 7/49 = 0.14$

$p(N|p) = 5/61 = 0.08$ $\quad p(N|n) = 42/49 = 0.86$

| | p | n |
|---|---|---|
| Y | 56 | 7 |
| N | 5 | 42 |

$$\text{Expected profit} = p(\boldsymbol{p}) \times [p(\boldsymbol{Y}|\boldsymbol{p}) \times b(\boldsymbol{Y}, \boldsymbol{p}) + p(\boldsymbol{N}|\boldsymbol{p}) \times b(\boldsymbol{N}, \boldsymbol{p})] + p(\boldsymbol{n}) \times [p(\boldsymbol{N}|\boldsymbol{n}) \times b(\boldsymbol{N}, \boldsymbol{n}) + p(\boldsymbol{Y}|\boldsymbol{n}) \times b(\boldsymbol{Y}, \boldsymbol{n})]$$

cost benefit

$$= 0.55 \times [0.92 \times b(\boldsymbol{Y}, \boldsymbol{p}) + 0.08 \times b(\boldsymbol{N}, \boldsymbol{p})] + 0.45 \times [0.86 \times b(\boldsymbol{N}, \boldsymbol{n}) + 0.14 \times b(\boldsymbol{Y}, \boldsymbol{n})]$$

$$= 0.55 \times [0.92 \times 99 + 0.08 \times 0] + 0.45 \times [0.86 \times 0 + 0.14 \times (-1)]$$

$$= 50.1 - 0.063 \approx \$50.04$$

The coefficient of determination, $R^2$ มีความเป็น linear มากน้อยขนาดไหน (0 - 1) ยิ่งเยอะ ยิ่งมีความเป็น linear เยอะ

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(observed_i - predicted_i)^2}{\sum_{i=1}^{N}(predicted_i - \overline{predicted})^2}$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(observed_i - predicted_i)^2$$

Mean Absolute Deviation (MAD)

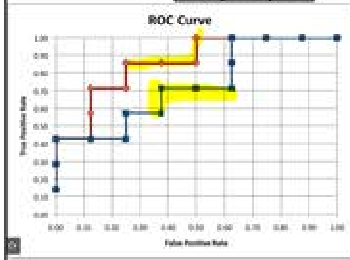
$$MAD = \frac{1}{N}\sum_{i=1}^{N}|observed_i - \overline{predicted}|$$

Roc มี Decision Threshold อยู่อะ

Decision Threshold $\theta = 0.5$

Predicted Class

| Spam | Non | Real Class |
|---|---|---|
| TP=4 | FN=1 | Spam |
| FP=3 | TN=2 | Non |




AUC = 0.50 ไม่ต่างอะไรกับการเดาสุ่มเลย

AUC > 0.70 คือเกณฑ์มาตรฐานสำหรับโมเดลส่วนใหญ่

AUC > 0.80 โมเดลทำงานได้ดี

AUC > 0.90 โมเดลทำงานได้ดีมาก

AUC เยอะแสดงว่าค่า threshold มีผลอย่างน้อย ยิ่งเยอะ ยิ่งดี

Performance Evaluate in test set

| Model | Accuracy | AUC |
|---|---|---|
| Classification Tree | 91.8%±0.0 | 0.614±0.014 |
| Logistic Regression | 93.0%±0.1 | 0.574±0.023 |
| *k*-Nearest Neighbors | 93.0%±0.0 | 0.537±0.015 |
| Naïve Bays | 76.5%±0.6 | 0.632±0.019 |

**Three-Way Hold-Out Strategy:** Divide the full dataset into three different subsets.

1. Training set: The subset of examples used for learning.
2. Validation set: The subset of examples used to tune the classifier (e.g. select parameter values).
3. Test set: The subset of examples used only to assess the performance of a fully-trained classifier.

**Classification**

1) **Hold-out sampling with validation set**
2) ***k*-Fold Cross Validation**
3) **Leave-one-out Cross Validation (Jackknifing)**
4) **Bootstrapping**

**Classification/Regression when a time dimension is concerned**

1) **Out-of-time Sampling**
2) **Walk-forward Validation**

**Leave one out:**Extreme case of k-Fold Cross Validation where k is selected to be the total number of examples in the dataset.
- For a dataset with n examples, perform n experiments.
- For each experiment use n-1 examples for training and the remaining single example for testing.
- Average the accuracy/error rates over all n experiments.
- In each fold, the test set contains only one instance.
- The training set contains the remainder of the data.

## Simple Linear Regression Analysis

การใช้ Regression Analysis ก็เพื่อต้องการหาสมการความสัมพันธ์ของตัวแปรเพื่อที่จะนำไปสู่การคาดการณ์หรือประมาณค่า ของตัวแปรที่เราไม่รู้ค่าโดยจะต้องมีการตรวจสอบเสียก่อนว่าสมการที่ได้มานั้นมีความถูกต้อง เพียงพอหรือไม่ แต่! ก่อนจะสร้างสมการควร visualize ดูก่อนเพราะบางครั้งสมการเหมือนกัน แต่ความสัมพันธ์ของข้อมูลต่างกัน

- **Linear regression, logistic regression,** and **support vector machines (SVM)** are all very similar instances of our basic fundamental technique:
  - The key difference is that each uses a different objective function

ตระกูลเดียวกัน แตกต่างกันที่ loss function ต่างกัน

**Logistic regression** is a class probability estimation model and not a regression model

Logistic regression is estimating the probability of class membership (a numeric quantity) over a categorical class

**SVMs** ก็คือเปลี่ยนจากหา loss function (minimize loss) ไปทำการให้ maximum margin
- Linear Discriminants, - Effective, - Use "hinge loss", Also nonlinear
- Hinge loss incurs no penalty for an example that is not on the wrong side of the margin

**loss function**
- **Zero-one loss** assigns a loss of zero for a correct decision and one for an incorrect decision
- **Squared error** specifies a loss proportional to the square of the distance from the boundary

Linear Model vs Tree Induction

- What is more comprehensible to the stakeholders?
  - Rules or a numeric function?
- How **"smooth"** is the underlying phenomenon being modeled?
  - Trees need a lot of data to approximate curved boundaries
- How **"non-linear"** is the underlying phenomenon being modeled?
  - If very, much "data engineering" needed to apply linear models
- **How much data do you have?!**
  - There is a key tradeoff between the complexity that can be modeled and the amount of training data available
- What are the characteristics of the data: missing values, types of variables, relationships between them, how many are irrelevant, etc.
  - Trees fairly robust to these complications

Linear Regression find the best line Y =B0 + B1X

bike rentals = 7501.8339 × *temperature* + 945.824

for every 1degree increase in the temperature, we would expect the bike rentals to increase by 7501 ::: B0 = intercept, B1 = slope

Use the least squares method to minimise the error(กำลังสองน้อยที่สุด)

With a little bit calculus we find the the values of $\beta_0$ and $\beta_1$ which minimise the squared residuals. These are **Ordinary Least Square** (OLS) estimates of $\beta_0$ and $\beta_1$

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{s_X^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - n(\bar{x})^2}$$

การหาความแตกต่างของค่าเฉลี่ย ระหว่างกลุ่มตัวอย่างตั้งแต่ 2 กลุ่มขึ้นไป

standard deviation of X $\quad$ covariance of X,Y

$$s_X = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \rightarrow Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$Cov(Y, X) > 0$ positive relationship

$Cov(Y, X) < 0$ negative relationship

$$Cor(X, Y) = \frac{Cov(X, Y)}{s_X s_Y}$$

Pearson Correlation: covariance scaled by the standard deviations

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{s_X^2} = \frac{199.221}{0.0266} = 7501.834$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 4504.348 - 7501.833 \times 0.4744 = 945.824$$

**SST** - the squared differences between the observed dependent variable and its mean.(total variable of the dataset) (sigma(yi - y-bar)^2)

**SSR** - sum of squares due to regression, or SSR. It is the sum of the differences between the predicted value and the mean of the dependent variable. (explain variability by your line) it means our regression model captures all the observed variability and is perfect. Once again, we have to mention that another common notation is ESS or explained sum of squares.