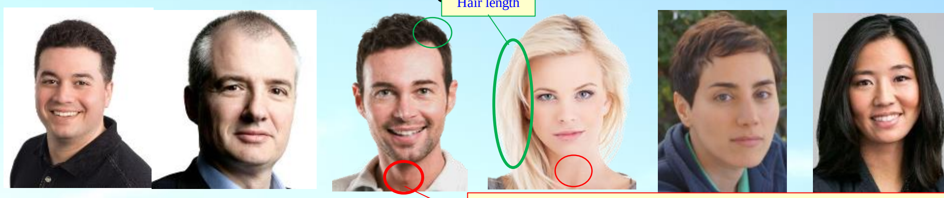
Chapter 2
Pattern Classification

Xuegong Zhang
September 16, 2021

2.1
Basic Concepts in Pattern Recognition

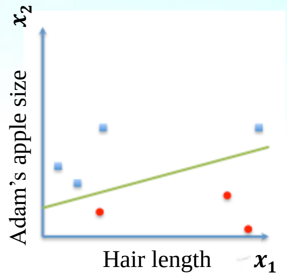
A toy example

- How to teach a machine to recognize man or woman?



The **Adam's apple**, or laryngeal prominence, is a feature of the human neck, and is the lump or protrusion that is formed by the angle of the thyroid cartilage surrounding the larynx seen especially in males.

A procedure of learning



For the green line:

$w_1x_1 + w_2x_2 + w_0 = 0$

For each blue dots:

$w_1x_1 + w_2x_2 + w_0 > 0$

For each red dots:

$w_1x_1 + w_2x_2 + w_0 < 0$

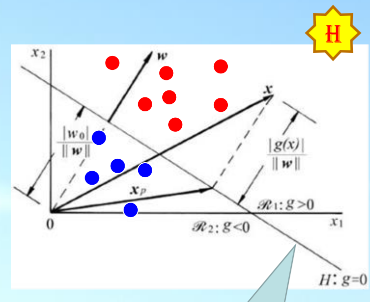
Linear classifier / Linear discriminant

$$y = f(x)$$

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

features

weights



$$y = \mathrm{sgn}\left(\sum_{i=1}^{d} w_i x_i + w_0\right)$$

$$y = \begin{cases} +1 & \Rightarrow \text{class } A \textit{ or } \mathbf{x} \in \omega_1 \\ -1 & \Rightarrow \text{class } B \textit{ or } \mathbf{x} \in \omega_2 \end{cases}$$

Class labels

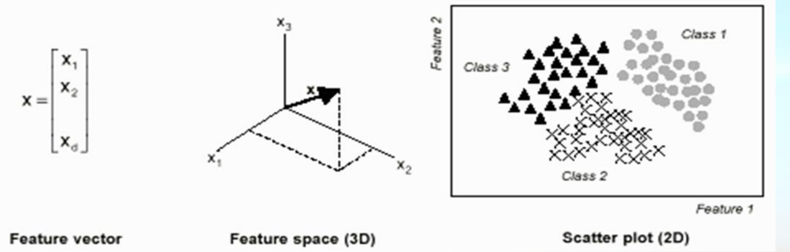
Decision line (boundary)

Terms and Concepts

- **Sample**
  - In statistics:
    - "A **set** of examples sampled from a population by a defined procedure"
    - Elements of a sample: **sample points**, or **instances**, or observations
    - **Sample size**: number of instances in a sample
  - In everyday life (and therefore often in CS):
    - "A small part or quantity intended to show what the whole is like"
    - Often (mis)used as equal to "**example**" or **instance**
      - "**A sample**" sometimes actually mean "one instance"
      - "**Sample set**" is used for a set of instances ("a sample")
      - "**Sample size**": number of samples in a sample set

Terms and Concepts

- **Features**
  - Quantitative attributes of a sample (that are relevant to the learning task)
  - Feature vector, Feature space
- **Class** (of samples)
  - A subset of samples that has some shared property → a region in the feature space



Feature vector   Feature space (3D)   Scatter plot (2D)

Terms and Concepts

- **Classifier** / Discriminant
  - A mathematical function of features that classify samples to classes

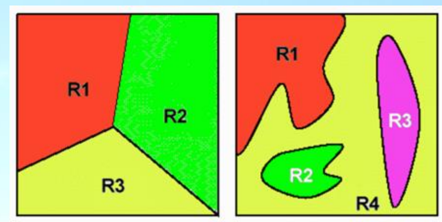
$$\mathbf{y} = f(\mathbf{x})$$

prediction ↑ classifier ↑ sample ↑

$$X = [\mathbf{x}_1, \cdots, \mathbf{x}_m] = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}$$

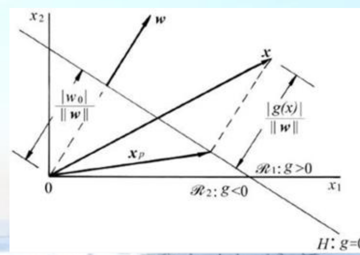
← features

A sample ↑



- Classification boundary / decision boundary
  - boundary between regions of classes in the feature space

- Linear discriminant / Linear machine: $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$
- Decision boundary: $g(\boldsymbol{x}) = 0$
- Decision rule: $g(x) \gtrless 0 \ \Rightarrow\ x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

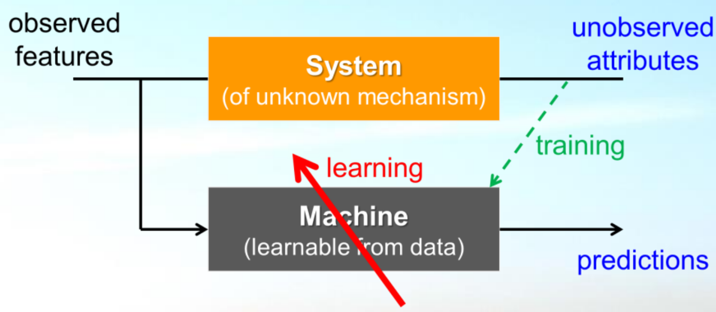


Terms and Concepts

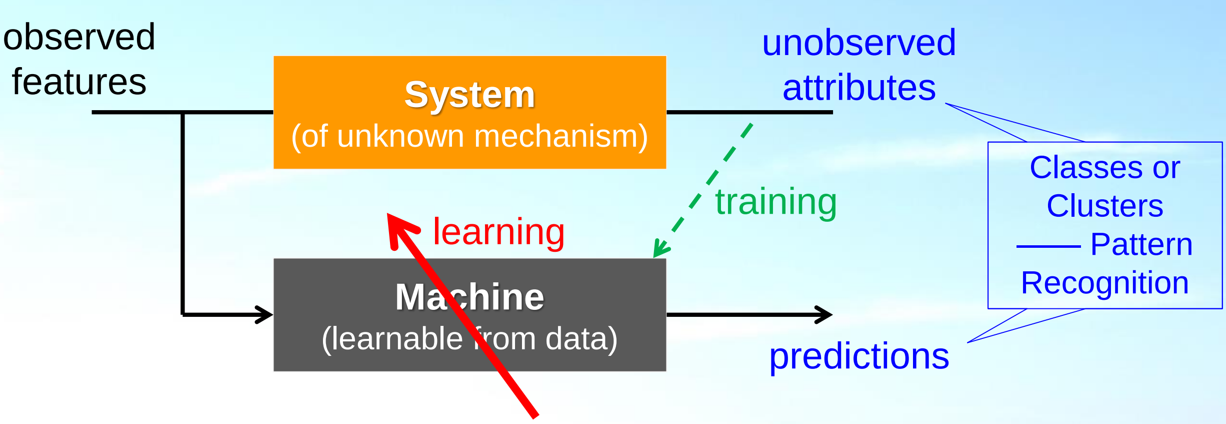
- Training
  - The procedure of fixing changeable parameters of a classification function according to data using some ML algorithm
- Training data
  - Samples with known classification labels for the ML algorithm to learn
- Unknown samples
  - Samples that we know their features but not their class labels
- Test samples
  - Samples that we pretend to be unknown, to be used to test learned machines

Terms and Concepts

- Supervised Learning / Supervised PR
  - Known samples are used for training the machine
- Unsupervised Learning / Unsupervised PR
  - Only unknown samples are used in the learning process



Machine Learning vs. Pattern Recognition



$$\boldsymbol{y} = f(\boldsymbol{x})$$

Machine Learning vs. Pattern Recognition

- PR is a type of tasks for ML (the most typical type)
- ML is the major type of methods for PR tasks

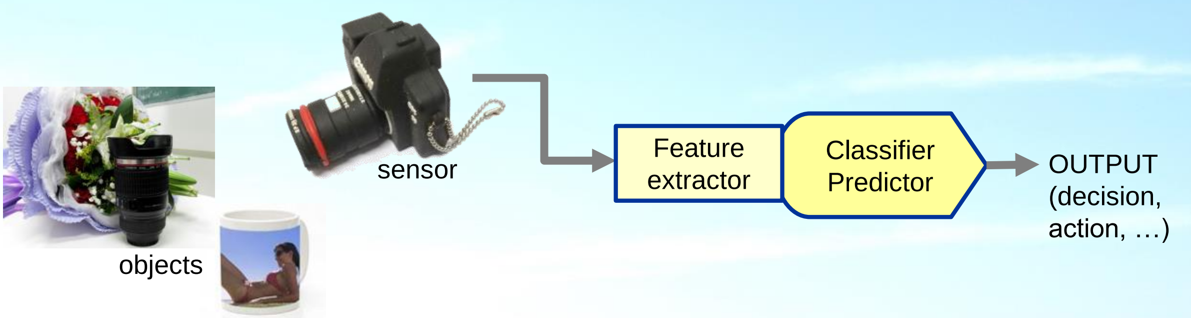
$$\underbrace{\mathbf{y}}_{\text{prediction}} = \underbrace{f}_{\text{classifier}}(\underbrace{\mathbf{x}}_{\text{sample}})$$

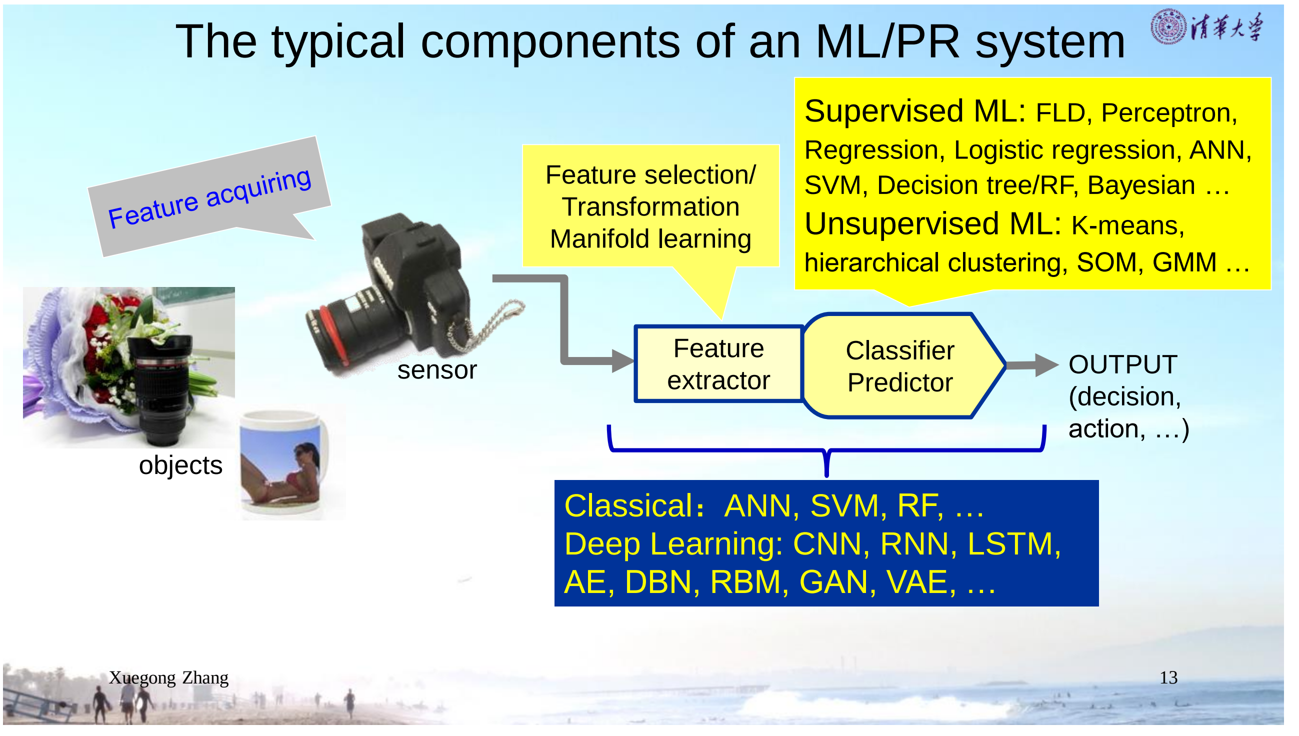
$$X = [\mathbf{x}_1, \cdots, \mathbf{x}_m] = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}$$

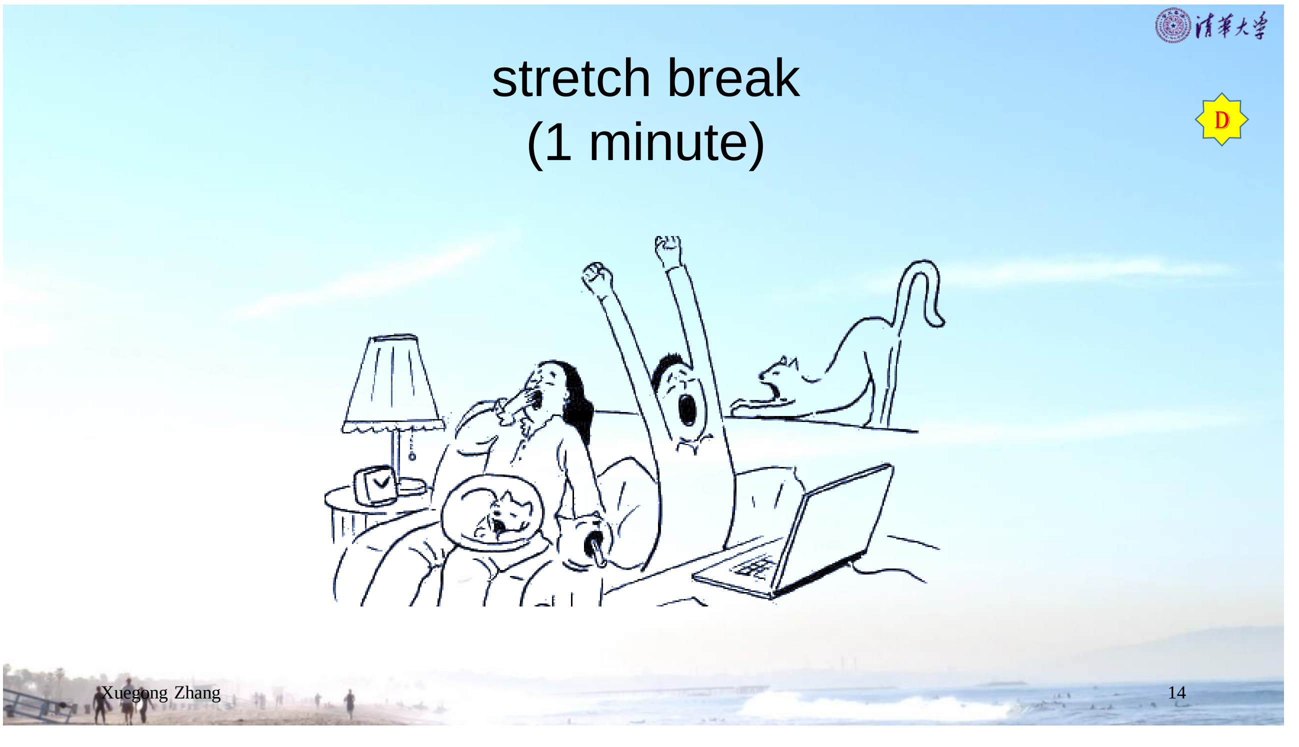
features

A sample

The typical components of an ML/PR system




The typical components of an ML/PR system
清華大學
Feature acquiring
objects
sensor
Feature selection/
Transformation
Manifold learning
Supervised ML: FLD, Perceptron, Regression, Logistic regression, ANN, SVM, Decision tree/RF, Bayesian …
Unsupervised ML: K-means, hierarchical clustering, SOM, GMM …
Feature extractor
Classifier Predictor
OUTPUT (decision, action, …)
Classical：ANN, SVM, RF, …
Deep Learning: CNN, RNN, LSTM, AE, DBN, RBM, GAN, VAE, …
Xuegong Zhang
13


清華大學
stretch break
(1 minute)
D
Xuegong Zhang
14

2.2
Assessment of Classifiers

2.2.1
Errors, Precision and Accuracy

Error and Error Rate

- An Error: a sample on which the classifier gives a wrong answer
- Error Rate:
  - Proportion of errors among all samples
    $$\varepsilon = \frac{k}{N}$$
    where k is the number of errors and N is the number of all samples.
  - Note: "error" is sometimes used for "error rate" in some literature.
- Are all errors equal?
  - Different mistakes may have different costs or consequences.

D

Terminology in diagnostic test

- Positive 阳性: something (e.g., a disease) is there
- Negative 阴性: the thing is not there
- Diagnostic Test: using a test to judge positive or negative → Classification
- Performance measures:
  - Sensitivity 灵敏度
    $$Sn = P(T^+ | D^+) = \frac{TP}{TP + FN}$$
  - Specificity 特异度
    $$Sp = P(T^- | D^-) = \frac{TN}{(TN + FP)}$$

	Disease present D^+	Disease absent D^-
Test positive T^+	TP: True Positives	FP: False Positives
Test negative T^-	FN: False Negatives	TN: True Negatives

Confusion matrix

Donald Trump (05/22/20): "yeah, I tested very positively in another sense. I tested positively toward negative, right. So now, I tested perfectly this morning."

• Terminology in Hypothesis Tests

- Hypothesis tests: null hypothesis vs. alternative hypothesis
- Two types of errors → Classification

Statistical decision	Real State	
	H0 true (N0) Negative	H1 True (N1) Positive
Reject H0 Declared positive	**Type I error** False Positive (**FP**) α (FP/N0)	Correct True Positive (**TP**) Sensitivity (TP/N1)
Accept H0 Declared negative	Correct True Negative (**TN**) Specificity (TN/N0)	**Type II error** False Negative (**FN**) β (FN/N1)

Performance measures for binary classification

- Type-I error (false positive rate): $\alpha = \frac{FP}{FP+TN}$
- Type-II error (false negative rate): $\beta = \frac{FN}{FN+TP}$
- Sensitivity (Sn, recall, detection rate): $Sn = \frac{TP}{TP+FN} = 1 - \beta$
- Specificity (Sp): $Sp = \frac{TN}{(TN + FP)} = 1 - \alpha$

	Disease present D^+	Disease absent D^-
Test positive T^+	TP: True Positives	FP: False Positives
Test negative T^-	FN: False Negatives	TN: True Negatives

- Precision (positive predictive value or PPV, discovery rate): $PPV = \frac{TP}{TP+FP}$
- Accuracy (Acc): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
- False discovery rate (FDR): $FDR = \frac{FP}{TP+FP}$
- F1 score (F-measure): $F_1 = \frac{2PPV \cdot TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$
- Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

	Disease present D^+	Disease absent D^-
Test positive T^+	TP: True Positives	FP: False Positives
Test negative T^-	FN: False Negatives	TN: True Negatives

More Jargons

Suggestion: always define the terms **in explicit forms** in your work and writing!

Terminology and derivations from a confusion matrix

condition positive (P)
the number of real positive cases in the data
condition negative (N)
the number of real negative cases in the data

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)
$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$
specificity, selectivity or true negative rate (TNR)
$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$
precision or positive predictive value (PPV)
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$
negative predictive value (NPV)
$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$
miss rate or false negative rate (FNR)
$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$
fall-out or false positive rate (FPR)
$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$
false discovery rate (FDR)
$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = 1 - \mathrm{PPV}$$
false omission rate (FOR)
$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TN}} = 1 - \mathrm{NPV}$$
accuracy (ACC)
$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

F1 score
is the harmonic mean of precision and sensitivity
$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$
Matthews correlation coefficient (MCC)
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$
Informedness or Bookmaker Informedness (BM)
$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$
Markedness (MK)
$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

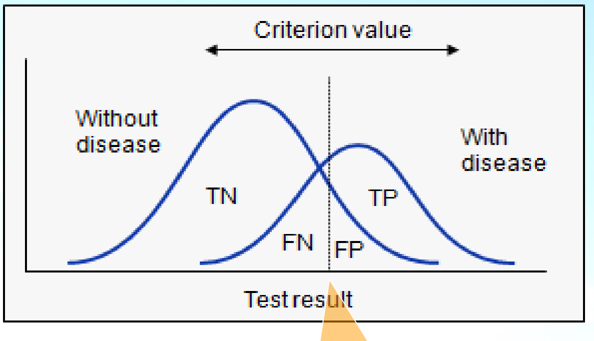
Sources: Fawcett (2006), Powers (2011), and Ting (2011) [2] [3] [4]

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

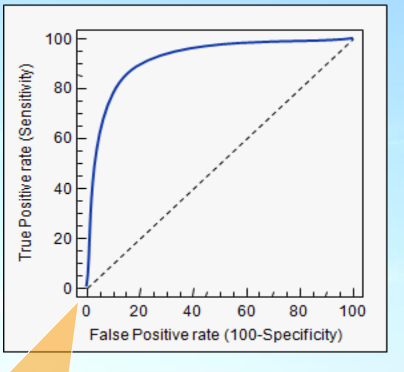
2.2.2
ROC Curves

Dependence of Test Results on the Threshold

- The trade-off between false positives and false negatives



adjusting the threshold

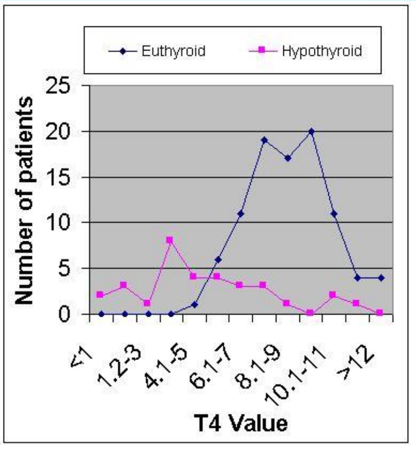


the threshold

the ROC curve

(https://www.medcalc.org/manual/roc-curves.php)

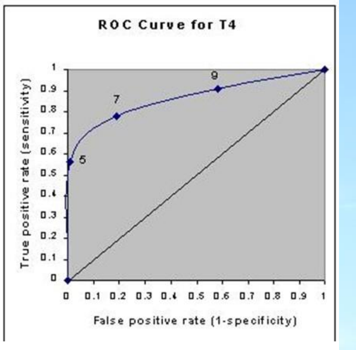
An example



Hypothyroid: 甲状腺机能减退
Euthyroid: 甲状腺机能正常

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cutpoint	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42



Thomas G. Tape, *Interpreting Diagnostic Tests*, Univ. of Nebraska Medical Center

Cutpoint	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

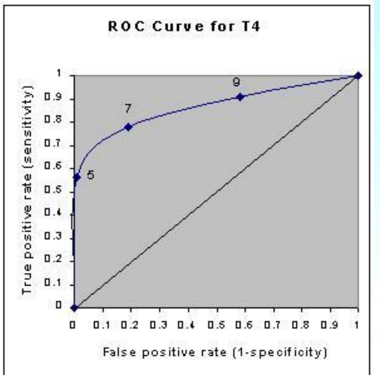
Thomas G. Tape, MD: *Interpreting Diagnostic Tests*

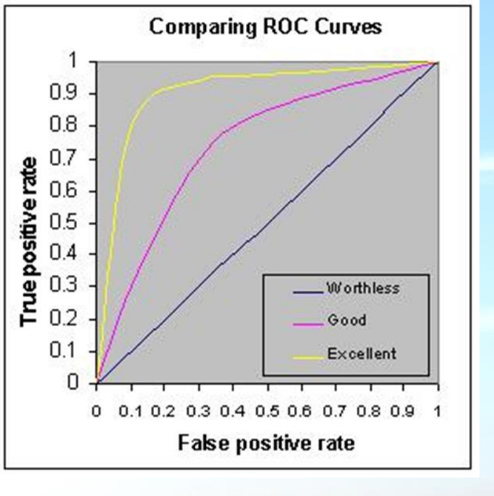
Xuegong Zhang 25

The ROC Curve: The Receiver Operating Characteristic curve

The name "Receiver Operating Characteristics" came from some study in signal detection theory during World War II. Came into the medical fields in 1970's.

- AUC: Area Under the Curve

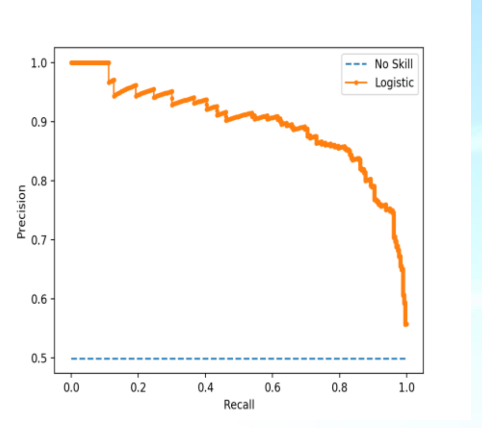




Xuegong Zhang 26

The PR Curve

- The Precision-Recall Curve



https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

单选题 1分 设置

4-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交

2.3 Experiment Design for Assessing Classifiers

2.3.1 Training and Testing

How to Estimate the Error Rate?

- Training Error: the error rate on the training set
  - Synonyms: *apparent error*, *re-substitution error*, *empirical error*
- True Error Rate: the expected error rate on future unknown data
- Independent Test Set (of size N), of known samples not seen in the training
- Test Error: the error rate on the test set

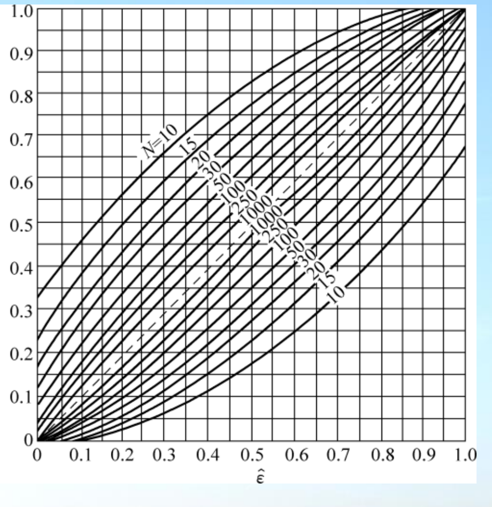
$$\hat{\varepsilon} = \frac{k}{N}$$

k is the number of wrong predictions on the test set.

Estimating the error rate

- Test error: the error rate on the test set
  $\hat{\varepsilon} = k/N$
- An estimate of the true error rate ε

- How good is the estimation?



Observation:

- Unbiased estimation, but the confidence of the estimation depends on the size of test set.

➔ A reliable estimation requires a large test set.

A dilemma when data are limited

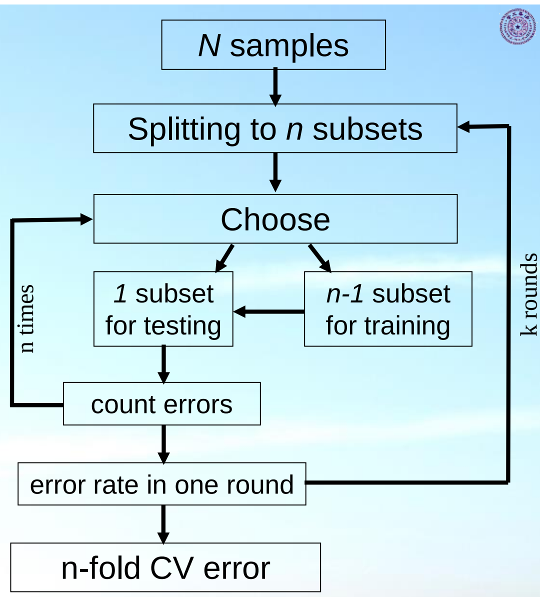
- For good estimation of the performance, we need a large test set
- But we want to use more data for the training

(why?) Let's leave it for future classes.

- Intuition:
  - More training data → better expected performance
    → smaller test set → less confidence in the assessment
  - Large test set → better assessment of performance
    → less training data → poorer expected performance

2.3.2 Cross Validation

Cross Validation (CV)



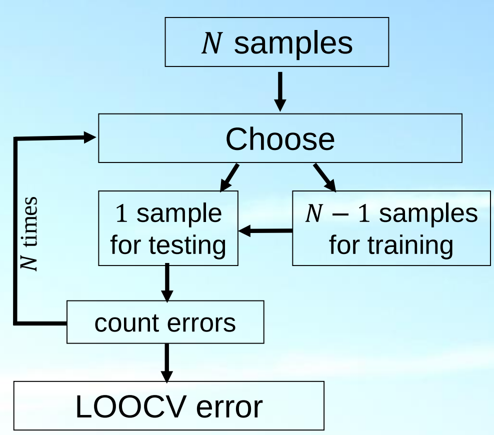
The n-fold cross validation

Typical:

- 3-fold CV
- 5-fold CV
- 10-fold CV

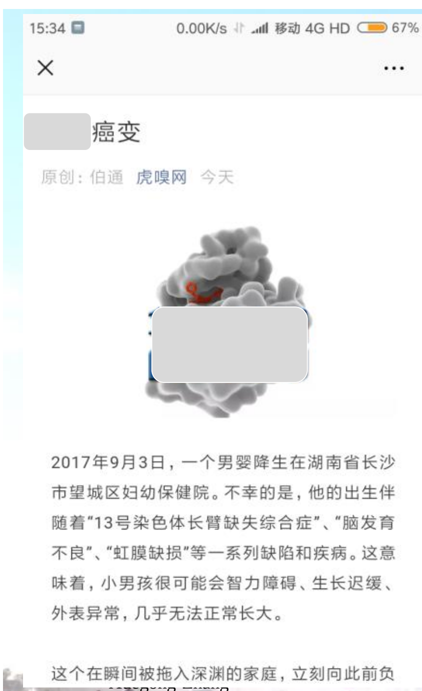
LOOCV

- The extreme case:
  Leave-one-out cross validation (LOOCV)
- Discussion
  - Widely used
  - Slightly conservative
  - Cross-validation provides almost unbiased estimation of the performance of a machine trained on $N-1$ or $\left(1-\frac{1}{n}\right)N$ samples.



2.4 Error Assessment in Real-World Situations

A real story

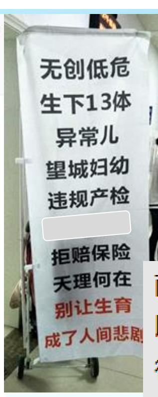

15:34 0.00K/s 移动 4G HD 67%

[redacted]癌变

原创：伯通 虎嗅网 今天

2017年9月3日，一个男婴降生在湖南省长沙市望城区妇幼保健院。不幸的是，他的出生伴随着“13号染色体长臂缺失综合症”、“脑发育不良”、“虹膜缺损”等一系列缺陷和疾病。这意味着，小男孩很可能会智力障碍、生长迟缓、外表异常，几乎无法正常长大。

这个在瞬间被拖入深渊的家庭，立刻向此前负



南方周末记者发现，早在2015年12月底，[redacted]基因便以无创产前基因检测准确率高达99.9%作为宣传点，进行全面推广。但在2018年7月13日晚间，[redacted]基因对外发布《[redacted]股份有限公司关于媒体报道的澄清公告》（下称《公告》）中这一数字悄悄发生了变化。

《公告》称，“国家卫计委发布《通知》 中，对于常规染色体非整倍体的检出率做了以下要求:21 三体、18 三体、13 三体的检出率分别不低于 95%、85%和 70%，而非 100%。 无创基因检测作为一项筛查技术存在假阴性的风险。”

A real story

- Company had claimed the accuracy of their genetic test for trisomy was 99.9
- A baby detected negative was born with trisomy 13 (Patau syndrome).
- Company said that MOH regulation required genetic tests for trisomy 13 to have detection rate > 70%, so they are innocent.

Accuracy 99.9%

Detection rate 70% (sensitivity)

南方周末记者发现，早在2015年12月底， 基因便以无创产前基因检测准确率高达99.9%作为宣传点，进行全面推广。但在2018年7月13日晚间， 基因对外发布《 股份有限公司关于媒体报道的澄清公告》（下称《公告》）中这一数字悄悄发生了变化。

《公告》称，"国家卫计委发布《通知》 中，对于常规染色体非整倍体的检出率做了以下要求:21 三体、18 三体、13 三体的检出率分别不低于 95%、85%和 70%，而非 100%。 无创基因检测作为一项筛查技术存在假阴性的风险。"

What was the problem?

D

- Company had claimed the accuracy of their genetic test for trisomy was 99.9%. (Acc=(TN+TP)/(N+P)=99.9%)
- A baby detected negative was born with trisomy 13 (Patau syndrome).
- Company said that MOH regulation required genetic tests for trisomy 13 to have detection rate > 70%, so they are innocent. (Sn=TP/P=TP/(TP+FN)>70%)

The Prevalence!

- Prevalence (*a priori* probability, prior probability)

$$Pre = P(D^+) = \frac{D^+}{D^+ + D^-}$$

	Disease present D^+	Disease absent D^-
Test positive T^+	TP: True Positives	FP: False Positives
Test negative T^-	FN: False Negatives	TN: True Negatives

- The prevalence of Patau syndrome is between 1 in 10,000 and 1 in 21,700 live births. (Wikipedia)

Hypothetical situations:

D

- If N+P=100,000, N=99,990, P=10, and
  - if TP=0, TN=99,990, FN=10, then Acc=99.99%, Sn=0%, Sp=100%
  - if TP=7, TN=99,990, FN=3, then Acc=99.997%, Sn=70%, Sp=100%

Take-home message?

- Accuracy may mean nothing if we don't consider the prevalence!

Xuegong Zhang 41

Another example: Shall I test my gene?

A hypothetic situation:

- There is a newly discovered genetics disease of prevalence of 10^{-6}. There is a test for the disease with sensitivity 100% and specificity 99.99%.

D

- Shall I take the test?
  - If I take the test and result is positive, what is the probability that I really have the disease?

Genetic Tests

Xue… 42

What we really care about

- False negative rate: $P(D^+|T^-)$
- True discovery rate: $P(D^+|T^+)$
- Bayes Formula (Bayes Rule):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(D^+|T^-) = \frac{P(T^-|D^+)P(D^+)}{P(T^-)} = \frac{(1-Sn)\,P(D^+)}{(1-Sn)P(D^+) + Sp \cdot (1-P(D^+))}$$

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)} = \frac{Sn \cdot P(D^+)}{Sn \cdot P(D^+) + (1-Sp)(1-P(D^+))}$$

Data in the two examples

- The Patau syndrome case:

$$P(D^+|T^-) \cong 0.3P(D^+) = 0.000003$$

assume $Sn = 0.7$ and $P(D^+) = 0.00001$

- The genetic test case:

$$P(D^+|T^+) \cong 1\% = 10000P(D^+)$$

Another question:

- How can the company estimate the sensitivity?

If N+P=100,000, N=99,990, P=10, and

if TP=0, TN=99,990, FN=10, then Acc=99.99%, Sn=0%, Sp=100%

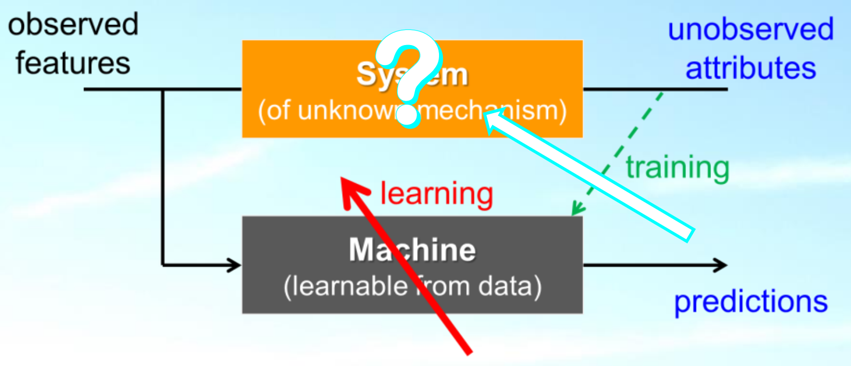
if TP=7, TN=99,990, FN=3, then Acc=99.997%, Sn=70%, Sp=100%

stretch break
(1 minute)

D

2.5
Inferring Relations based on Classification Performances

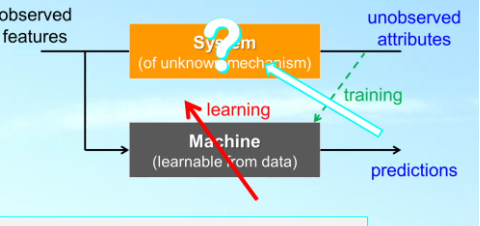
Using PR/MR for Scientific Discovery?



Q: If the machine predicted the attributes with certain accuracy, can we infer that the system exists?

The Principle of Scientific Inference

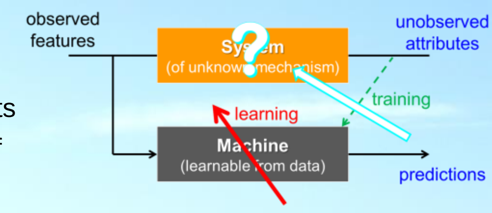
- Basic principle of data-based scientific inference
  - Generate a reasonable hypothesis
  - Assume the hypothesis is wrong and observe what happens (under the null hypothesis)
  - Try hard to reject the hypothesis
  - **If failed, accept it as a scientific discovery**

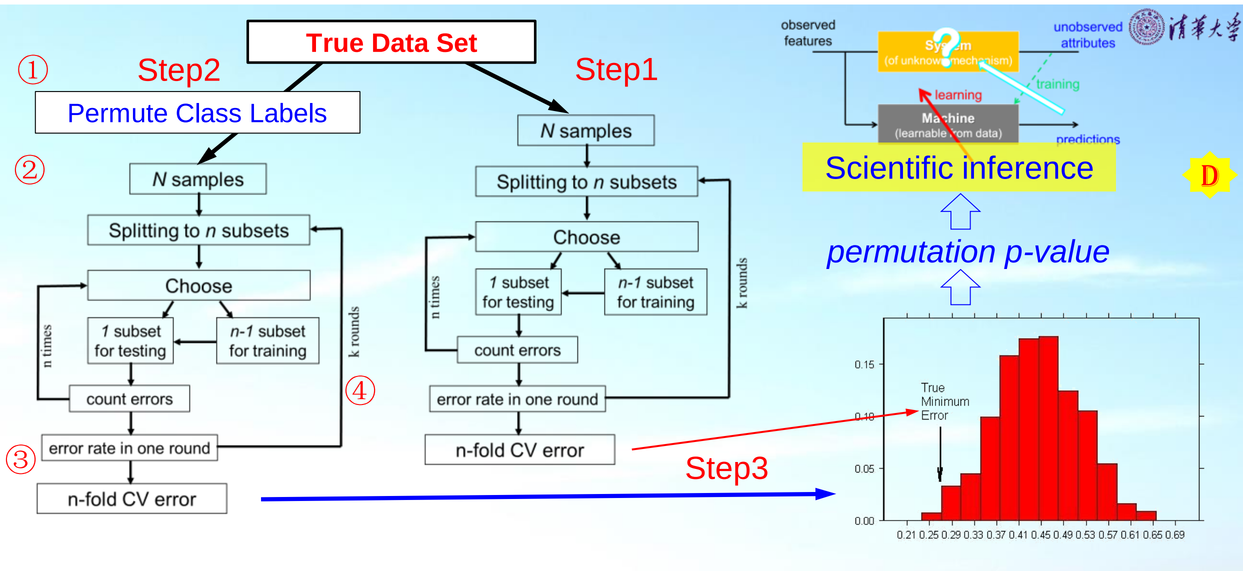


e.g. *p-value* < 0.05 (a psychological threshold)
- The probability of getting the observation under the null hypothesis.
- "If the hypothesis is false, the chance of getting the observation is less than 5%. So I choose to believe it is true."
- "I choose to make this claim since the chance of a false claim (type-I error) is 5%".

Inferring Relations with Permutation Tests

- Permutation Test for ML performances
  1. Apply the machine on the data as is
  2. Apply the machine on permuted data:
     ① Permute the data by shuffling the expected outputs
     ② Let the machine learns on the permuted data as if they are real
     ③ Observe the performances on permuted data
     ④ Repeat ① ~ ③ for a predefined number of rounds
  3. Compare the performance on true data with that on permuted data ➔ p-value





2.6 Discussion

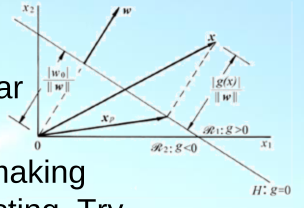
Two scenarios of ML studies

- Competition Scenario
  - Task pre-defined
  - Data given
  - Assessment method standardized
  - Goal: to show the I'm better
- Real-world Scenario
  - Task to be defined
  - Data to be collected
  - Assessment method to be designed
  - Goal: to solve the real-world problem

Being aware of what we are doing!

Homework

- Problems (Pr1)
  1. Basic concepts of ML.
  2. Basics exercise on linear discriminants.
  3. A hypothetic decision-making problem in Covid-19 testing. Try to answer it from the viewpoint of an individual and of the public.



D

- Deadline:
  – Sept. 22, 23:00 Beijing Time
- Computer exercises (Ex0)
  – (Optional) A basic task with Python. Get yourself prepared for future exercises.
  – Nothing to hand in.
  – TAs may give a tutor session on Python if >20 students emailed the need by Sept.18.
- Deadline:
  – Sept. 29, 23:00 Beijing Time

Xuegong Zhang 53