Machine Learning

< Block I : Regression >.

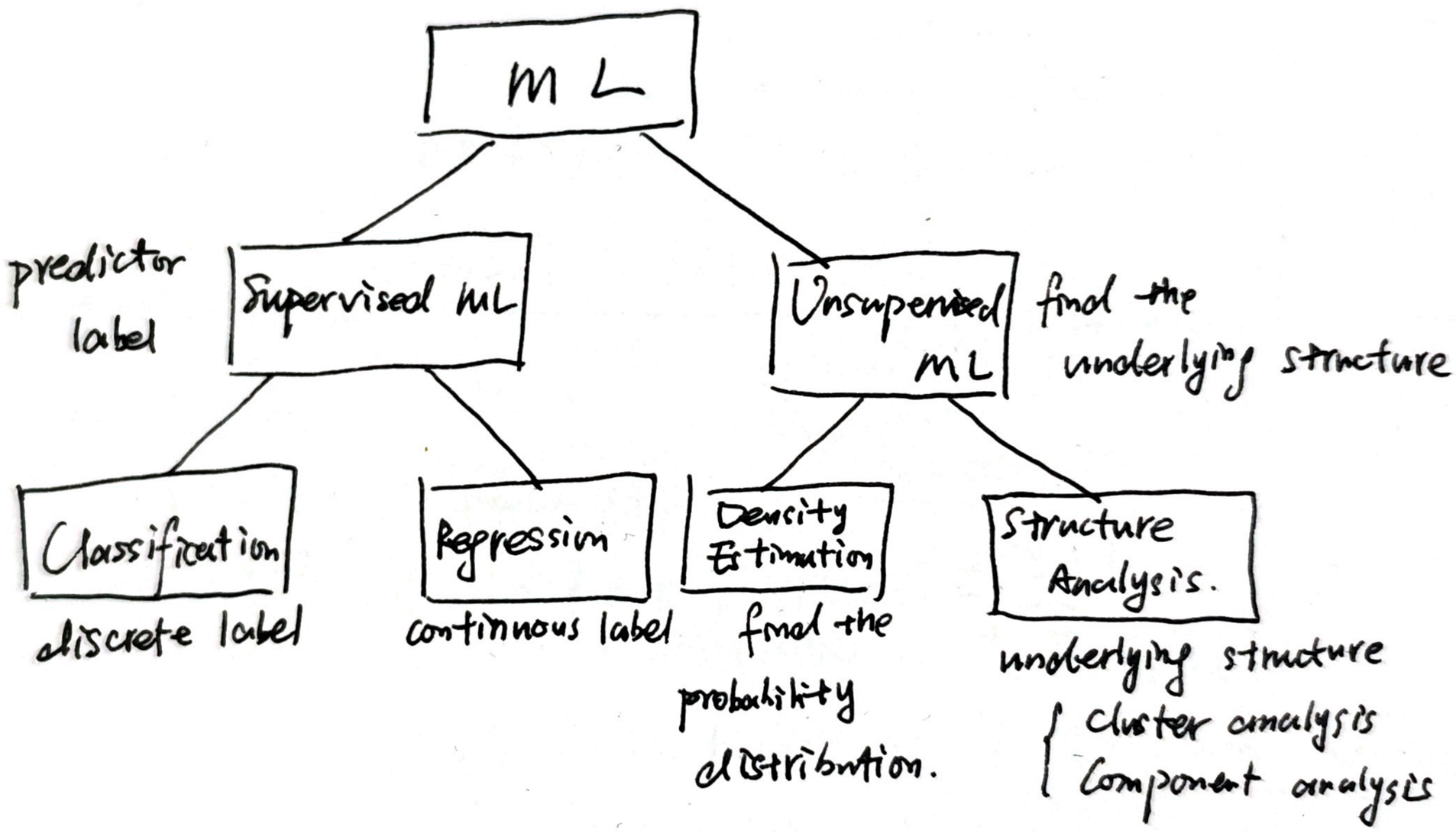
0. Introduction.

ML: The ability to acquire knowledge, by extracting patterns from data.

Data : Instance of an observation or a measurement.

Datasets : Data formatted as collections of items described by a set of pre-defined attributes.

ML : A set of tools together with a ~~method~~ methodology for solving problems using data.



1. Regression.

Define metric. Let $y = \hat{y} + e$, Sum of squared errors : $SSE = \sum e_i^2$

mean squared error : $MSE = \frac{1}{N}\sum e_i^2$

Root mean squared error : $RMSE = \sqrt{\frac{1}{N}\sum e_i^2}$

mean absolute error : $MAE = \frac{1}{N}\sum |e_i|$

R-squared : $E_R = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$

Model multiple linear regression

$Xw = Y$.

$$\begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,k} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

每行代表一个样本，使用最小二乘

$\min \frac{1}{N}\|Xw - Y\|^2$ 可解得

$w = (X^T X)^{-1} X^T Y$.

Remark 多项回归的高次形态也会备拟合, 拿一个 predictor 举例. $XW=Y$.

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^k \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

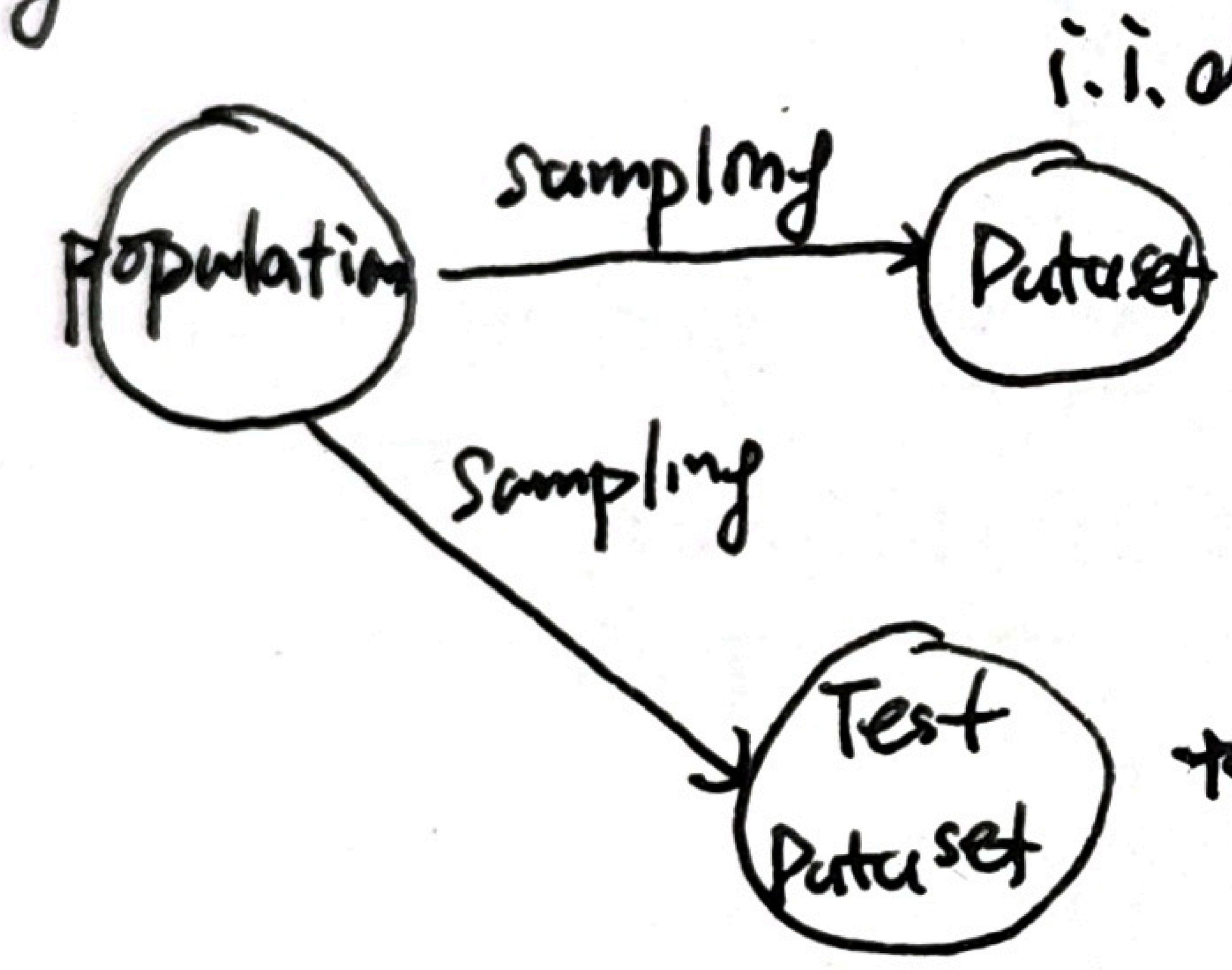
, 求解方式依然 $W=(X^TX)^{-1}X^TY$.

Define 泛化. Generalisation is the ability of our model to successfully translate what we was learnt during learning stage to deployment. (translate knowledge)

Define 泛化能力
1) Underfitting: unable to describe the underlying pattern. Large train and deployment errors are produced.
2) Overfitting: memorisation of irrelevant details. Small errors in train, large errors in deployment.
过拟合, 模型复杂, 会把噪声拟入.
3) Just right: Reflects underlying pattern and ignores irrelevant details.

2. Methodology I.

model
target population.



Dataset: empirical description of target population.

Test Dataset: test deployment performance as an estimation of true performance.

Algorithm GD

Gradient descent: $w_{t+1} = w_t - \varepsilon \nabla \mathcal{L}(w)$, update iteratively where ε is called learning rate or step size.

Remark
1) 并不是在 error surface 上优化, 而是在 Train dataset 描述的 empirical error surface 上优化.
2) 实际使用时, 将 Training data 分为多个 Batch, 每次求 ∇ 时选择 N 个 subset (Batch) 进行梯度的求解 → Stochastic gradient descent.
原因: small batches produce noisy of gradient of empirical error surface, which can help to escape local minima.
3) other 改进: momentum; RMSProp ↦ Adam.

Regularisation. Modifies the empirical error surface by adding a term that constrains the values that the model parameters can take on. (ex. $\frac{1}{N}\|XW-Y\|_2^2+\lambda\|W\|_2^2$)

Remark 使用 Regularisation 可以减少 model 的 complexity, 在训练的时候使用的应当称之为 cost or objective function, 是带正则化项的. 但是 deployment 时用的指标, 是 target quality metric, 不带正则化项, Regularisation 只是为了在训练时作用.

Validation Validation allow us to use data for assessing and selecting different families of models, then the one will be trained.

1) Validation set approach Randomly splits the available dataset into training and a validation (holdout) set.

2) Leave-one-out cross-validation (LOOCV). The validation set contains only one sample. Mulitple split, final performance is calculated as the average of individual performances.

3) k-fold cross-validation 分成k个group，最后取平均

Remark 这些所有的 cross-validation 都是为了得到 validation error，以此来直接得到也可以去平均多次得到，但是都是在得到了各个误差，我们需要对每选择的 model 或超参都运行一遍 cross-validation，得到每个model 对应的 validation error，再比较进行 "model selection"，随后将模型在全部训练集和验证集上再训练一遍 model 作为最终 output model.

总结：最终的 mode quality 取决于 type of model, optimisation strategy and the representativeity of the training data.

<Block 2: Classification>

1. Classification 1: KNN 相关.

problem Classification.

1) 以 attribute space 表示 dataset，label 作为一个 axis 来把每个点标出. 以 predictor space 表示 dataset, using different symbols for each label in the predictor space.

2) A solution model in classification is a partition of the predictor space into decision regions 和 separated by decision boundaries.

Define 线性分类

In linearly separable datasets, we can find a linear classifier that achieves the maximum ~~accuracy~~ accuracy ($A=1, E=1-A=0$). In non linearly-separable datasets, the best accuracy will be $A<1$, error rate $E>0$.

Algorithm

logistic regression.

一个训练线性分类器的策略 strategy

Let $P(\text{class } 1) = p(w^T x_i) = \dfrac{e^{w^T x_i}}{1+e^{w^T x_i}} = \dfrac{1}{1+e^{-w^T x_i}}$, Then use MLE. Thus.

$$\max_w L = \max_w \prod_{\substack{\text{class 2} \\ (j)}} (1-p(w^T x_j)) \cdot \prod_{\substack{\text{class 2} \\ (i)}} p(w^T x_i)$$

or.

$$\max_w l = \max_w \log L = \max_w \sum_{\substack{\text{class 2} \\ (j)}} \log[1-p(w^T x_j)] + \sum_{\substack{\text{class 1} \\ (i)}} \log p(w^T x_i).$$

Remark 1) logistic Regression 可以看作一种进行 Linear classifier 的准则：keep that Boundary away，即最大似然，只有远离边界，才可最大似然。

2) 可以令 $d_i = w^T x_i$，d_i 即 point i 距离 Boundary 的欧式距离，则可以用 $p(d_i)$ 计算整体的 likelihood.

Algorithm

Nearest neighbours

Non-parametric approaches：K-Nearest Neighbours (KNN)

Instance-based method：The whole training dataset need to be memorised.

给定一个point，寻找最近的K个point，返回归类为K个point中多类的类别 (most popular class).

K增加，boundary become less complex，overfitting ⟵(less K) (larger K)⟶ underfitting.

总结：1) classifiers are <u>partitions</u> of the <u>predictor space</u> into <u>decision regions</u> <u>separated by boundaries</u>.

2. Classification II：概率视角.

model

Bayes classifier

$$P(\text{class}\,k|x) = \frac{P(x|\text{class}\,k)\,P(\text{class}\,k)}{P(x)} \propto p(x|\text{class}\,k)\,P(\text{class}\,k)$$

$p(x|\text{class}\,k)$：likelihood, class densities；$P(\text{class}\,k)$：Prior.

对于单纯的 Bayes classifier. $\frac{P(\text{class}\,1|x)}{P(\text{class}\,2|x)} \gtrless 1$ → Class1 / Class2 → 考虑 Risk，看具体情况.

Use $\frac{P(\text{class}\,1|x)}{P(\text{class}\,2|x)} \geq T$ 去做判别，使用 true posterior probabilities 的分类器为 Bayes classifier.

→ 考虑 Bayesian Risk (min the expected cost). cost → $\frac{C_1 \times P(\text{class}\,1|x)}{C_2 \times P(\text{class}\,2|x)} \geq 1$.

Discriminant analysis

Algorithm

一种使用参数估计 prior 和 Class densities 的方法，从而得到 posterior.

对 likelihood (class densities) 的假设是 Gaussian

$$P(x|\text{class}\,i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}$$

对先验的估计是频率诱导的概率.

Note 1) 对于 Discriminant analysis for 2分类，$\Sigma_1 = \Sigma_2$（初始假设），boundary 是线性的 → Linear DA, LDA.

如果 $\Sigma_1 \neq \Sigma_2$，boundary 是二次的，Quadratic DA, QDA.

2) 使用 Discriminant analysis、logistic Regression、KNN 都是对 posterior 的估计，在 Bayesian Risk 中，通过调整 C_1, C_2 (Risk)，可以用 Risk 或概率的意义上指导 boundary 的选取.

Metrics

Confusion matrix

对于 Class-sensitive problem 和 imbalanced dataset，使用：$Cost1 \times P(\text{class}\,1|x) = Cost2 \times P(\text{class}\,2|x)$

Accuracy 和 error rate 可能并不是很好的选择，可以考虑使用 Confusion Matrix.

predicted class \ Actual class	P	N
$\hat{P}$	TP	FP
$\hat{N}$	FN	TN

FP：错误的分为 $\hat{P}$

⟼

	P	N
$\hat{P}$	TPR	FPR
$\hat{N}$	FNR	TNR

TPR + FNR = 1.

FPR + TNR = 1.

Note 1) 通过 Confuse matrix 还可以衍生出一些指标：

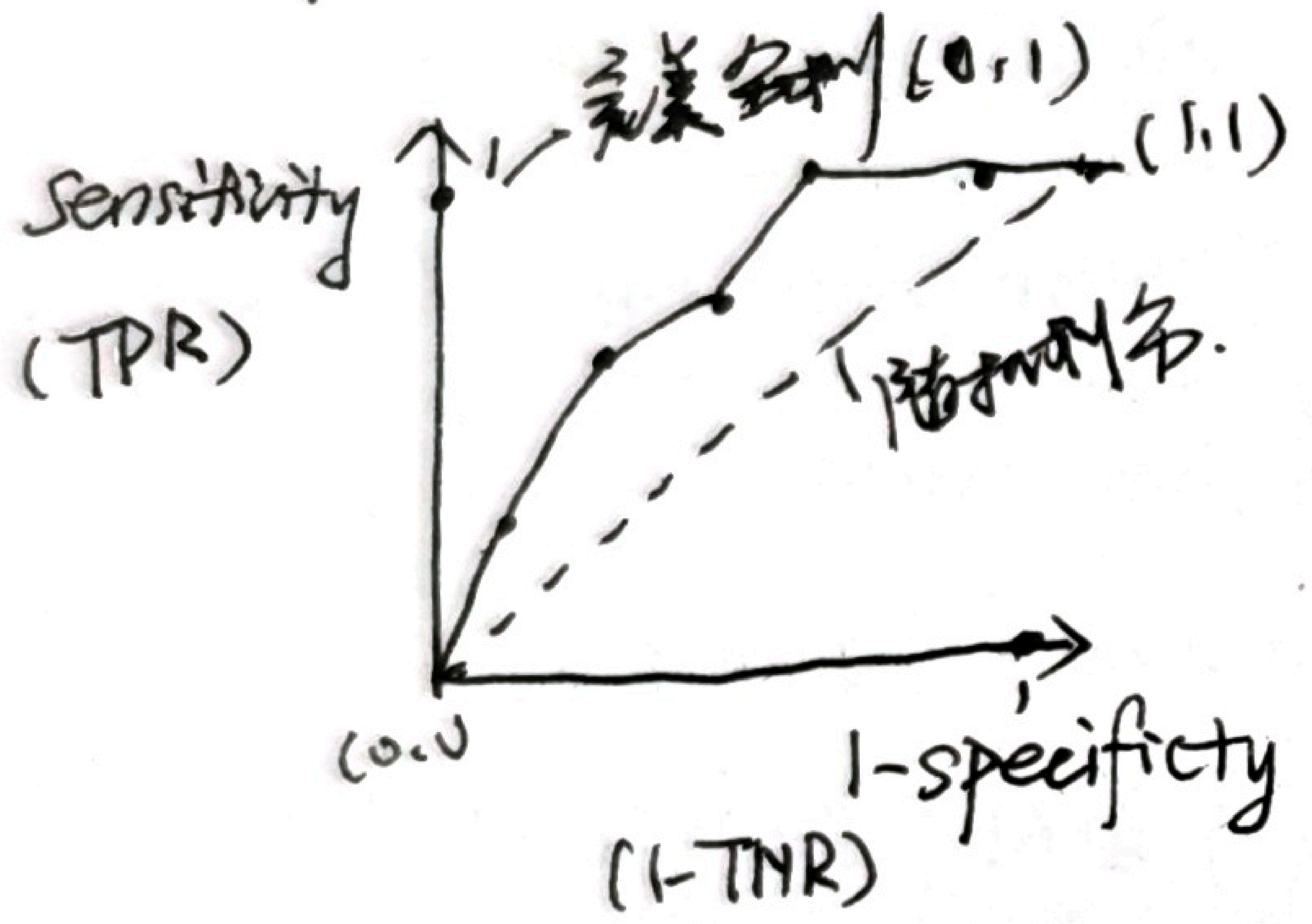
- Sensitivity (recall / true positive rate) = TPR. 灵敏度，召回率.
- Specificity (true negative rate) = TNR 特异性
- Precision (positive predictive value) = $\frac{TP}{TP+FP}$ 精度，P中正确的概率.

2) 优化单个指标会带来问题，when improve one metric, others may be descrease. 所以有时会考虑 pairs of quality metrics.

ex. F1 score：$F1 = \frac{2}{recall^{-1} + precision^{-1}} = 2\frac{Percision \times recall}{Percision + recall}$.

3) 前文中提到了 Risk Dicision，$\frac{P(class1|x)}{P(class2|x)} \geq T$，如何进行T的选择呢？

可以通过 Confuse matrix 来完成这一点，但一堆矩阵是很难得到直观理解的，可以画出 ROC (Receiver operating characteristic) plane，ROC是 TPR 关于 (1-TNR) 的曲线，反映了二者的变化情况. allow us to explore class-sensitive quality metrics.



① 选取不同的T.
② 根据T生成 decision boundary，得到 Confuse matrix.
③ 将 TPR和(1-TNR)画入ROC连线

ROC体现了所有Confuse matrix的 TPR和TNR关系！方便来做好的对比 matrix.

最后选择的T是距离 (0,1) 最近的点对应的T.

4) 现在我们可以对比单个model了，如何对比不同model呢？ ↦ **AUC!**

AUC — Area under the curve，即ROC曲线下的面积

注：AUC 也可以从概率视角进行解释，它反映了model捕捉到 map 的能力.
该 map 指 score of P > score of N 的映射，这就是数据的 latent variable.

好的AUC → 1，差的AUC → 0.5，[illegible]，根据AUC可进行 model selection.

5) 对AUC的一句评价：AUC is a measure of goodness for a classifier that select T. → can be calibrated 校准.
不仅是T，也指model的 parameter

<Block 3：Methodology, DL, NNs>.

1. Methodology II.

a. pipeline: The term pipeline is often used to describe workflows (sequence of operations). ML solution are more than a model it include severl stages form a processing pipeline: Transformation stages, Severl ML ~~models~~ models, Aggregation ~~stages~~ stage.
所有的 pipeline 都涉及 ~~deploym~~ deployment.

b. Normalisation Data normalisation is a tunable transformation stage that allow us to scale attributes so that their value belong to similar range.

1) Min-max normalisation.

$$z = \frac{x - min(x)}{max(x) - min(x)} \in [0, 1]$$，使得特征有[0,1]范围.

2) Standardisation.

$z = \frac{x-\mu}{\sigma}$, 且 $z=0$, 且 $z^2 = 1$., ensures inputs are treated equally.

Note $min(x)$, $max(x)$, μ, σ 都是 train dataset 中获得的 parameter, 于 test data 或, 区整合及也在 deployment 在 test data 中, 但是可能会出现 out-of-range values, 甚至 outliers

c. Trasformations Transformations are data manipulations that change the way that we represent our samples ⇒ move the data to another space.

类型: 同dim space 变换, 不同 dim space 变换 (dimensionality reduction)

PCA Principal components analysis.

1) identifies the directions along which samples are aligned. These directions define a destination space with the same number of dim as the original space, i.e.
build a linear transformation and additionally assigns a score to each component.

2) We can use the score (eigenvalue) to Rank the attributes, define the destination space and Remove the least important attributes.
that is, use PCA to reduce dimension.

Dimensionality Reduct

1) Feature selection. Assume that only a subset of the original attributes are relevant.

① Filtering. Consider each attribute individually, don't consider their interactions.

② Wrapping. Consider the interaction between features, evaluate them together.

注: 使用 Filtering 和 Wrapping 的方法一般: 使用不同 subset 训练, validation validation, rank, 最终的 validation performance 对应的 subsets.

2) Feature extraction. New attributes are defined as operations on the original attributes.

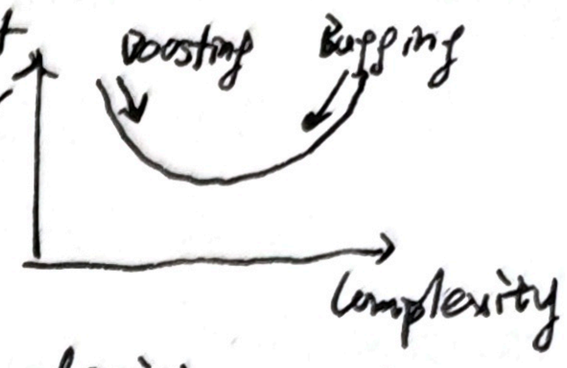
Note Transformation 同样也为 learn! selecting the right transformation (via validation) tuning the parameters of given transformation (via training) 6

d. Ensembles. Combines the strengths of base models.

1) Bagging. Generates K sub-dataset by bootstrapping and train K simple base models with each-dataset. $f(x) = \frac{\sum f_{k-100}}{K}$

where Bootstrap is a statistical method that extracts random samples from a dataset. (Randomisation).

2) Boosting. Generates a sequence of simple base models, where each successive model focus on the samples that the previous models could not handle properly.



Note Bagging 是并行算法, Boosting 依赖前一个 model, 串行.

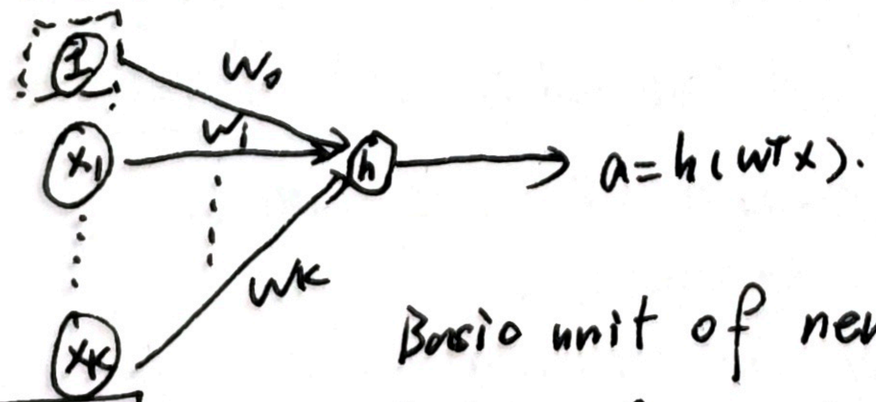
Decision trees. Partition the predictor space into multiple decision regions by implementing sequences of splitting rules using only one predictor only. (每次用一个 predictor, 各个的 Region boundary 是 axis-parallel 的).

Random Forests. Train many individual trees by randomising the training samples and the predictors. Predictions are obtained by averaging the individual predictions. ↦ Bagging.

2. Neural Networks and Deep Learning.

Patterns are laws in our data, structure is a law in our target population, we should discovering the underlying structure by identifying the pattern.

perception. Defined by a weight vector w and an activation function $h(x)$, that map an extended x to the output 2. w_0 is bias.



Basic unit of neural network!

h → step activation function → linear classifier

h → logistic function → logistic Regression.

Common ↦ grid pattern detector.

Remark. 1) A layer is a collection of perceptrons that use the same input.
ex. L 个 perceptrons, K 个 input, 有 $(K+1) \times L$ 个 weight.

2) Architecture. describe how layers are Connected.
hidden layers produce internal features.

3) A neural network consists of perceptrons arranged in layers that are connected arr according to an architecture.

NNs as ML models. NNs can be seen as an entire tunable ML pipeline.

① Cognitive point. large NNs is flexibilty to create new and increasingly complex concepts that might be relevant to predict.

② ~~ML model 自有~~ ML Algorithm 自有 model, cost function 和 optimisation method.
对于NNs 来说，可以用 Back-propagation 来有效地计算 Gradient, start from output.

③ Transfer learning. 训练好的模型不变, fixed transformation stage T(x)
再 retrain late stage using new data f(z), (z=T(x)).
transferring an already learnt transformation.

Types of layers.

1) Fully-Connected layer (FC).

2) Convolutional layer.

问题：Images and ~~time series~~ 都表现出 equivariance property：
Same pattern can be expected in different locations of the grid.

Solution: Perceptrons 组排列为 grid, 称为 feature map.
share their parameter → Convolutional kernel!
卷积层由多个 feature maps 组成，每个由不同 kernel 处理，different Concept, 最后 stack 在一起形成 Convolutional layer.

Note. ~~dimension~~ dimensions of kernel are H×W×D; 画面积, weights per kernel is H×W×D+1(bias), Train 卷积层，计算时训练的 kernel 的 H×W×D+1 个 weight. → kernel 中的权值 +bias.

3) Pooling layers.

Reduces the size of feature maps. 由多个邻域压缩成一个的考虑，插入 Convolutional layers 之间.

① Max pooling ② Average pooling 无须训练!

总结：1) 从 Functional (map), Cognitive (create new concepts), Computational 角度看 NNs.
2) NNs 不是 ML model, ~~是算法~~ 而是当作生成算法的算法. 即 NNs 都为一系列 architecture, 不同的 architecture 是不同 model. (通过 validation 模型选择).

<Block4: Unsupervised learning>.

1. Structure analysis.

Unsupervised learning: Understanding how samples are distributed in the attribute space.

method1: Structure analysis: identify regions within attribute space (cluster) or directions with high density (component).

method2: Density estimation: quantify the probability that find the sample within a region in attribute space.

应用: Discover structure, generate new knowledge, change the way we represent the data.

Algorithm — k-means.

Clusters: Describle the structure of dataset as group.

定义两种 proximity-based quality metric:

① intra-cluster sample scatter: $I(C_i)=\frac{1}{2}\sum_{x_i,x_j \in C_i}||x_i-x_j||_2^2$

② inter-cluster sample scatter: $O(C_0,C_1)=\sum_{x_i \in C_0, x_j \in C_1}||x_i-x_j||_2^2$.

Best clustering: lowest intra-cluster and highest inter-cluster.

补整 ---

实际上减少 intra 就是增加 inter

k-means 用 center (mean) of cluster 表示 prototypes: $\mu=\frac{1}{N}\sum_{x_i} x_i$.

上述 intra-cluster sample scatter 有另一种等价计算: $I=N_0\sum_i||x_i-\mu||_2^2$.

So that good clustering: samples are close to their prototype

算法: ① Prototypes are obtained as the center of each cluster.

② Samples are re-assigned to the cluster with the ~~closte~~ closest prototype.

Note 1. ~~超~~ Hyperparameter K 的选择:

① 实际问题

② validation, 但不是一般的验证误差小, 不准确, 因为当 $K=N$, $Err=0$.

使用 elbow method, when $K>K_T$, The improvement of quality will be slower.

Note 2. k-means 只可 produces spherical clusters. (Convex-set), 不可处理 Non-Convex 情况.

使用 density-based clustering 可以解决.

Algorithm — DBSCAN. Density-based spatial clustering of applications with noise.

两参数, radius r, threshold t, 可将所有样本分为3类.

Core: 一个 r 内, 样本个数大于等于 t.

Border: 一个 r 内有 core 点, 但是 core 点小于 t.

Outlier: others.

core点邻域中的 core点为一个 cluster (实际上所 core 名为 cluster, 领邻域内合并).
Border 归属于邻域内最多 core点 归属的 cluster. Outlier 不归属.

<u>Note</u> DBSCAN 不需设置 Number of class, 实际上它可以对 sample 自动 Cluster

Algorithm
层次 cluster

<u>Hierarchical clustering</u>. The structure of a dataset can be <u>explored</u> at i.e. progressively building <u>different levels</u> that expose <u>different properties</u>.

{a, b, c, d} — whole dataset

↙ ↘

{a, b} {c, d}

↙ ↘ ↙ ↘

{a} {b} {c} {d}. each sample cluster

↑ top … ↓ bottom

divisive top-down ↓ splits. ↑ merges agglomerative bottom-up.

两种 build dendrogram 的方法.

dendrogram 系统树图

Relationship between clusters at different levels.

<u>Note</u> merge or split 的 option:

① Single linkage $D(A,B) = \min\{d(x,y) : x \in A, y \in B\}$ → arbitrary shape

② Complete linkage $D(A,B) = \sum_{x\in A, y\in B} d(x,y)$.

③ Group average: $D(A,B) = \sum_{x\in A, y \in B} d(x,y) / |A||B|$

(② ③ → spherical shape.)

<u>补充</u> Component analysis allows us to identify the <u>directions in the space</u> that our <u>data are aligned with</u> (PCA).
应用: transform data, clean it, reduce dimensionality.

2. Density Estimation.

<u>Assumption</u> Samples distributed in the attribute space, And this distribution can ~~be~~ be ~~repersent~~ repersent or described by <u>Probability densities</u>. (视为model).

<u>Note</u> 这里的 model (Probability density) 指 true distribution of data, 我们需要用 ML model 和 data 去 Density Estmation.

<u>Algorithm</u>
Non-parametric.

Do not specify the <u>shape</u> of probability.

1) Histogram. $\hat{f}(x) = \frac{1}{Nh}$ (Number of x_i in same [bin] as x) (bin → h.)

b → small ⇒ spiky, b → big ⇒ flat. (Not flexible to find the underlying structure).

2) kernel density estimation. (KDE). Let $h \to 0$, And $Nh \to \infty$.

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right).$$

Building an <u>individual density</u> around each sample and combining they.

Algorithm

parametic method

Gmm.

Parametric approaches specify the shape of the probability density.

Gaussian Mixture models (GMM). mixing coefficients.

$$p(x) = \sum_m \pi_m N(x|\mu_m, \Sigma_m), \text{ where } \sum_m \pi_m = 1.$$

Note 1) Why Gaussian? ①CLT ②由1,2阶矩决定. ③ independent ⇔不相关

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

2) Gaussian 的参数 parameters estimation (unbias).

$$\hat{\mu} = \frac{1}{N}\sum x_i, \quad \hat{\Sigma} = \frac{1}{N-1}\sum (x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

最大似然(ML)定义下, $\hat{\Sigma} = \frac{1}{N}\sum (x_i-\hat{\mu})(x_i-\hat{\mu})^T$.

3) Algorithm: Expectation-Maximization (EM).

E-step: 计算 latent variable posterior, 得到 Expected likelihood over fix θ.

M-step: $\arg\max_\theta$ likelihood (latent variable, θ).

Outliers samples just depend on noise ⇒ outliers or anomalies.

1) Anomaly detection algorithm quantify the probability, if $p(x_i) < T$, x_i anomaly.

2) 在设计算法时, 为了 mitigate outliers 的影响, 增加算法的 robustness 的目的.

App for classifiers 1) LDA, QDA, use Bayes to learn.

2) Naive Bayes., Assume that predictor are independent

$$h_{Bayes} = \arg\max_{y\in\{0,1\}} P(Y=y)P(X=x|Y=y).$$

$$= \arg\max_{y\in\{0,1\}} P(Y=y)\prod_i^d P(X_i=x_i|Y=y).$$

将 $d+d^2$ 的参数减少到 $d+d$ (只含对角元素). ⇒ 降维了.

App for Cluster ① K-means as a version of Gmm with $\Sigma = \sigma^2 I$.

If $\Sigma = \sigma^2 I$, Gmm produce spherical decision region.

No.

Date

Machine learning.

1. 最小范数，最小二乘、mp逆.

通用的，存在 $A \in \mathbb{R}^{m \times n}$，存在问题

$$\min_x \|Ax - b\|_2^2$$

该问题为在对于行满秩和列满秩的情况... 一行空间.

Solution. $A = U\Sigma V^T = [U_R \quad \underbrace{U_N}_{\text{左零空间}}] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix}$ 零空间.

$\underbrace{\quad}_{\text{列空间}}$

$$\min_x \|Ax - b\|_2^2 = \min_x \|U^T Ax - U^T b\|_2^2$$

$$= \min_x \left\| \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix} x - \begin{bmatrix} U_R^T \\ U_N^T \end{bmatrix} b \right\|_2^2$$

$$= \min_x \|\Sigma_r V_R^T x - U_R^T b\|_2^2 + \|U_N^T b\|_2^2$$

$$= \min_x \|U_R \Sigma_r V_R^T x - U_R U_R^T b\|_2^2 + \|U_N^T b\|_2^2$$

$$= \min_x \|Ax - U_R U_R^T b\|_2^2 + \|U_N^T b\|_2^2$$

(proj(A) = $AA^+ = U_R U_R^T$; proj(A^T) = $A^+A = V_R V_R^T$ 见<补充>.)

把 b 投影到 A 列空间 $x^* = \arg\min_x \|Ax - U_R U_R^T b\|_2^2 = V_R \Sigma_r^{-1} U_R^T b \triangleq A^+ b$.

此时, $\min_x \|Ax - b\|_2^2 = \|U_N^T b\|_2^2$.

对应的最优解为 $x = A^+ b + v$, $v \in N(A)$. 两

<含A行满秩>. ▭ ↦ 最小范数.

此时, $m < n$, 方程组欠定, 无数组解. 需对其刻化.

$$\min \|x\|_2^2$$

$$s.t. \quad Ax = b.$$

Analysis. 上述求解 $\|x\|_2^2 = \|A^+ b\|_2^2 + \|v\|_2^2 \geq \|A^+ b\|_2^2$

即 $\min \|x\|_2^2 = \|A^+ b\|_2^2$.

验证. 可以证明, 最小范数解 $x = A^T (AA^T)^{-1} b$

此时 $A = U_R [\Sigma_r \quad 0] \begin{bmatrix} V_R^T \\ V_N^T \end{bmatrix}$.

代回计算有 $x = A^T (AA^T)^{-1} b = A^+ b$.

四

No.

Date

2. 最可能的就是最对的！最大似然！

最大似然估计（Maximum Likelihood Estimation, MLE）.

Assumption：强假设，目标分布已知，只是参数未知（θ未知）.

Motivation 使用分布 Q w.r.t. $\hat{\theta}$ to approximate 分布 P w.r.t. θ 真实分布.

（频率学派视角：任何分布都有确定的（单值）参数！）.

Solution. 运用 KL divergence 来度量 P 和 Q.

$$D_{KL}(P_\theta(x) \| P_{\hat{\theta}}(x)) \quad \text{（同分布，参数不同，} \hat{\theta} \text{为估计参数）}$$

$$= E_{P_\theta} \log \frac{P_\theta(x)}{P_{\hat{\theta}}(x)}$$

$$= E_{P_\theta} \log P_\theta(x) - E_{P_\theta} \log P_{\hat{\theta}}(x).$$

由大数定理. $E_{P_\theta} \log P_{\hat{\theta}}(x) \Rightarrow \frac{1}{N} \sum \log P_{\hat{\theta}}(x_i)$

$$D_{KL}(P_\theta(x) \| P_{\hat{\theta}}(x)) = E_{P_\theta} \log P_\theta(x) - \frac{1}{N} \sum \log P_{\hat{\theta}}(x_i).$$

（等号右侧）θ 必然有最优值，$E_{P_\theta} \log P_\theta(x)$ 为 Constant.

$$\hat{\theta}^* = \arg\min_{\hat{\theta}} D_{KL}(P_\theta(x) \| P_{\hat{\theta}}(x))$$

$$= \arg\min_{\hat{\theta}} -\frac{1}{N} \sum_{i=1}^{N} \log P_{\hat{\theta}}(x_i)$$

$$= \arg\max_{\hat{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log P_{\hat{\theta}}(x_i)$$

$$= \arg\max_{\hat{\theta}} \frac{1}{N} \log \prod_{i=1}^{N} P_{\hat{\theta}}(x_i).$$

$$= \arg\max_{\hat{\theta}} \prod_{i=1}^{n} P_{\hat{\theta}}(x_i) \Rightarrow \text{已知的 } P_{\hat{\theta}}(x_i), \text{指的是 } P_r(x_i | \hat{\theta}).$$

口.

最终，我们得到 $\hat{\theta}^* = \arg\max_{\hat{\theta}} \prod_{i=1}^{N} P_{\hat{\theta}}(x_i)$

$$= \arg\max_{\hat{\theta}} \text{likelihood}(\hat{\theta})$$

Remark. 最后的 object 就是"最大似然"，直观理解其思想：

当 Sample 点够多时，我取到的样本 $D = \{(x_i, y_i), i \in N\}$ 就是最可能取到的 dataset.

缺点：（假设多）（但仅仅对 Probability 建模而言）.

小样本大数定理不适用，MLE 会偏离.

KOKUYO

No.

Date

Remark 2 对似然函数的讨论不论到此为止，

Define likelihood function $\mathcal{L}(\theta|x) = p(x|\theta) = P_\theta(x)$.

$\arg\max_\theta P_\theta(x) = \arg\max_\theta p(x|\theta) = \arg\max_\theta \mathcal{L}(\theta|x)$,

所以，$P_\theta(x)$，$P(x|\theta)$，$\mathcal{L}(\theta|x)$ 都可称为似然，，频率学派中，写作 $P(x;\theta)$，$P(x,\theta)$ 的居多，因为不把 θ 当 r.v. 来看待.

考虑 Bayes formula：$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{P(x)}$.

（$p(x|\theta)$ → 似然，$p(\theta)$ → 先验）

3、2分类中的最优分类器.

对任何 $S = ((x_1,y_1), \dots)$ where $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ 定义 loss.

$$R(h) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{f_h(x) \neq y}].$$

其中，h 为假设集内的一种 function，2分类器 $f_h(x)$ 为

$$f_h(x) = \begin{cases} 1, & h(x) \geq 0 \\ -1, & h(x) < 0 \end{cases}.$$

Define $y(x) = Pr(y = +1 \mid X = x)$

则有 $R(h) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{f_h(x) \neq y}]$

$$= \mathbb{E}_{x\sim D}[y(x)\mathbb{1}_{h(x)<0} + (1-y(x))\mathbb{1}_{h(x)\geq 0}]$$

Define Bayes optimal classifier $h^*(x) = y(x) - \frac{1}{2}$.

即，为某个label的概率超过 1/2，选它！

Lemma $\forall h \in \mathcal{H}$, s.t. $R(h) - R(h^*) = 2\mathbb{E}_{x\sim D^x}[|h^*(x)|\mathbb{1}_{h(x)h^*(x) \leq 0}]$

proof. $R(h) = \mathbb{E}_{x\sim D^x}[y(x)\mathbb{1}_{h(x)<0} + (1-y(x))\mathbb{1}_{h(x)\geq 0}]$

$$= \mathbb{E}_{x\sim D^x}[y(x)\mathbb{1}_{h(x)<0} + (1-y(x))(1-\mathbb{1}_{h(x)<0})]$$

$$= \mathbb{E}_{x\sim D^x}[(2y-1)\mathbb{1}_{h(x)<0} + (1-y(x))]$$

$$= \mathbb{E}_{x\sim D^x}[2h^*(x)\mathbb{1}_{h(x)<0} + (1-y(x))]$$

$$\text{对于 } R(h) - R[h^*]$$

$$= \mathop{\mathbb{E}}_{x\sim D^x}\left[\, 2h^*(x)\left(\mathbb{1}_{h(x)<0} - \mathbb{1}_{h^*(x)<0}\right)\right].$$

$h(x)$	$h^*(x)$	$\mathbb{1}_{h(x)<0} - \mathbb{1}_{h^*(x)<0}$	
1	1	0	
-1	-1	0	
1	-1	−	$\Rightarrow -h^*(x) = \lvert h^*(x)\rvert$.
-1	1	+	$\Rightarrow +h^*(x) = \lvert h^*(x)\rvert$.

$$\text{对于 } R(h) - R(h^*)$$

$$= 2\mathop{\mathbb{E}}_{x\sim D^x}\left[\, \lvert h^*(x)\rvert\, \mathbb{1}_{h(x)h^*(x)\le 0}\right]. \quad \text{Q.E.D.}$$

□.

Remark 该 Lemma 表示了, Bayes classifier 是最优分类器.
其实对于所有(不B.Z分类), Bayes 也是最优.

$$f_{\mathcal{D}}(x) = \begin{cases} 1, & P(y=1 \mid x=x) > \frac{1}{2} \\ 0\text{-}1, & \text{otherwise.} \end{cases}$$

4. Linear Discriminant Analysis [Intr to ML. Lecture 6].

Logistic Regression 模型对 $P(\text{class} \mid \text{data})$

Discriminant Analysis 模型对 $P(\text{data} \mid \text{class})$.

A. GDA.

DA 模型对 likelihood $P(\text{data} \mid \text{class})$ 做出假设 $f_k(x) \sim N(\mu_k, \Sigma_k)$

由贝叶斯公式:

$$P(\text{class}=k \mid x) = \frac{P(x \mid \text{class}=k)\, P(\text{class}=k)}{\sum_i P(x \mid \text{class}=i)\, P(\text{class}=i)}$$

$$\propto \frac{f_k(x)\,\pi_k}{P(x)} \propto f_k(x)\,\pi_k.$$

现在我们假设处理二分类问题, 且 $\pi_{k1} = \pi_{k2}$ ($\pi_1 = \pi_2 = \frac{1}{2}$).

使用 MAP 准则获取 decision boundary

$$P(\text{class } 1 \mid x) = P(\text{class } 2 \mid x)$$

$$\Rightarrow P(x \mid \text{class } 1)\,\pi_1 = P(x \mid \text{class } 2)\,\pi_2$$

$P(x|\text{class }1) = P(x|\text{class }2) \Leftrightarrow \log f_1(x) = \log f_2(x)$. ①

$f_k(x) \sim N(\mu_k, \Sigma_k) \Rightarrow f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\}$.

对①式整理，有

$$\log|\Sigma_1| + (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) = \log|\Sigma_2| + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$$

这是由二次型表达的边界，所以称为 QDA.

B. LDA.

对QDA model 增加一条假设： $\Sigma_1 = \Sigma_2 = \Sigma$.

则①式可以整理为：

$$X^T\Sigma^{-1}(\mu_1-\mu_2) = \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^T\mu_2) = \text{Const}.$$

即把 x 在 $\Sigma^{-1}(\mu_1-\mu_2)$ 方向投影.

Σ^{-1} 可以起到纠正 $(\mu_1-\mu_2)$ 的作用，如果没有 Σ^{-1} 的作用，或 $\Sigma = \sigma^2 I$，那么分离超平面垂直于质心中垂线，LDA 退化为 nearest centroid classifier

再看，若 $\Sigma = \sigma^2 I$，$\pi_{k1} \neq \pi_{k2}$，则分离边界以垂直于有质心的连接的方式在连线上移动.

C. Estimating the parameters. (π_k, μ_k, Σ_k).

QDA：

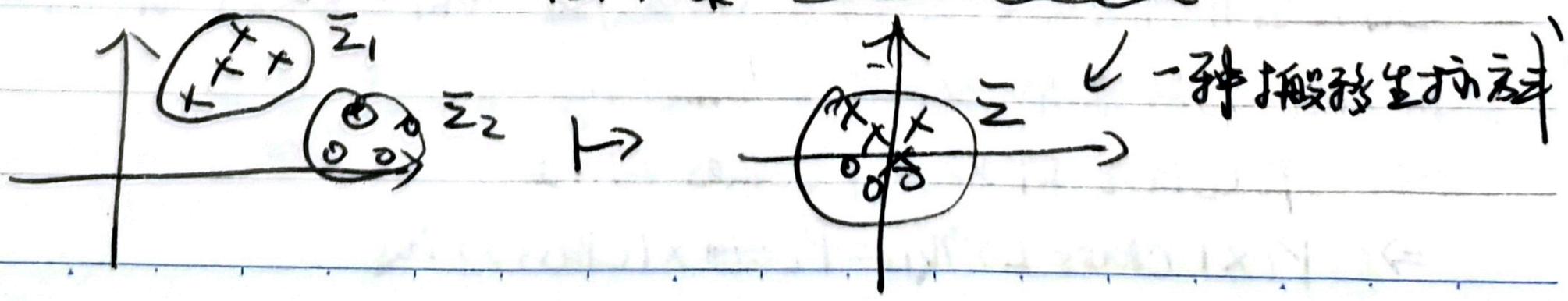
$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i\in C_k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k-1}\sum_{i\in C_k}(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}.$$

LDA：pooled covariance estimator：

$$\hat{\Sigma} = \sum_{k=1}^{K}\frac{1}{n-K}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\mu_k)(x_i-\mu_k)^T$$



No.

Date

D. model Choose.

Σ^{-1} 求逆运算会有精度问题，$\Sigma^{-1}(\mu_1-\mu_2)$ 带来 high-variance overfitting 的问题，常常使用 Regularization 解决.

Recall: $(X^TX)^{-1} \mapsto (X^TX+\lambda I)^{-1}$

$\Rightarrow \Sigma^{-1}(\mu_1-\mu_2) \mapsto (\Sigma+\lambda I)^{-1}(\mu_1-\mu_2)$

$\Rightarrow ((1-\lambda)\Sigma+\lambda I)^{-1}$, $\lambda \in [0,1]$.

也叫 mix QDA & LDA.

$\Sigma^{-1} \mapsto ((1-\lambda)\Sigma_k+\lambda\Sigma)^{-1}$.

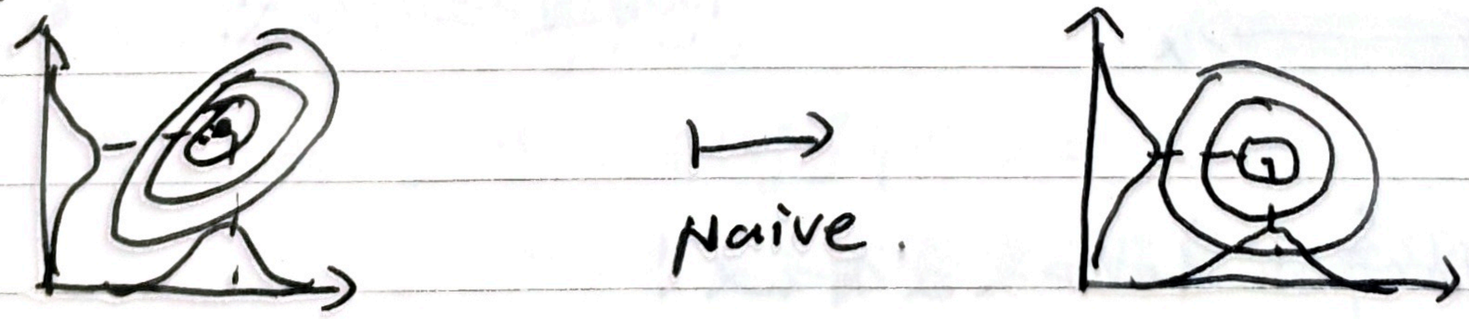
Covariance matrics	Separate	shared
Full	QDA	LDA
Diagonal	Naive Bayes	Diagonal LDA
Spherical	'Spherical QDA'	Nearst Centroid

注: Diagonal QDA aka Gaussian Naive Bayes.

Assume $f_K(x) = \prod_{j=1}^{p} f_{kj}(x_j)$

去除各个特征相关性! (Naive).

$f_k \sim N(\mu_k, \Sigma_k)$, $\Sigma_k = diag\{\sigma_{k1}^2 \cdots \sigma_{kp}^2\}$, $f_{kj} \sim N(\mu_{kj}, \sigma_{kj}^2)$



Naive.

E. Fisher Discriminant Analysis.

LDA 的另外的推导.

Problem: Find a linear projection that would maximize the radio of the between-class 'spread' to the within-class 'spread'.

KOKUYO

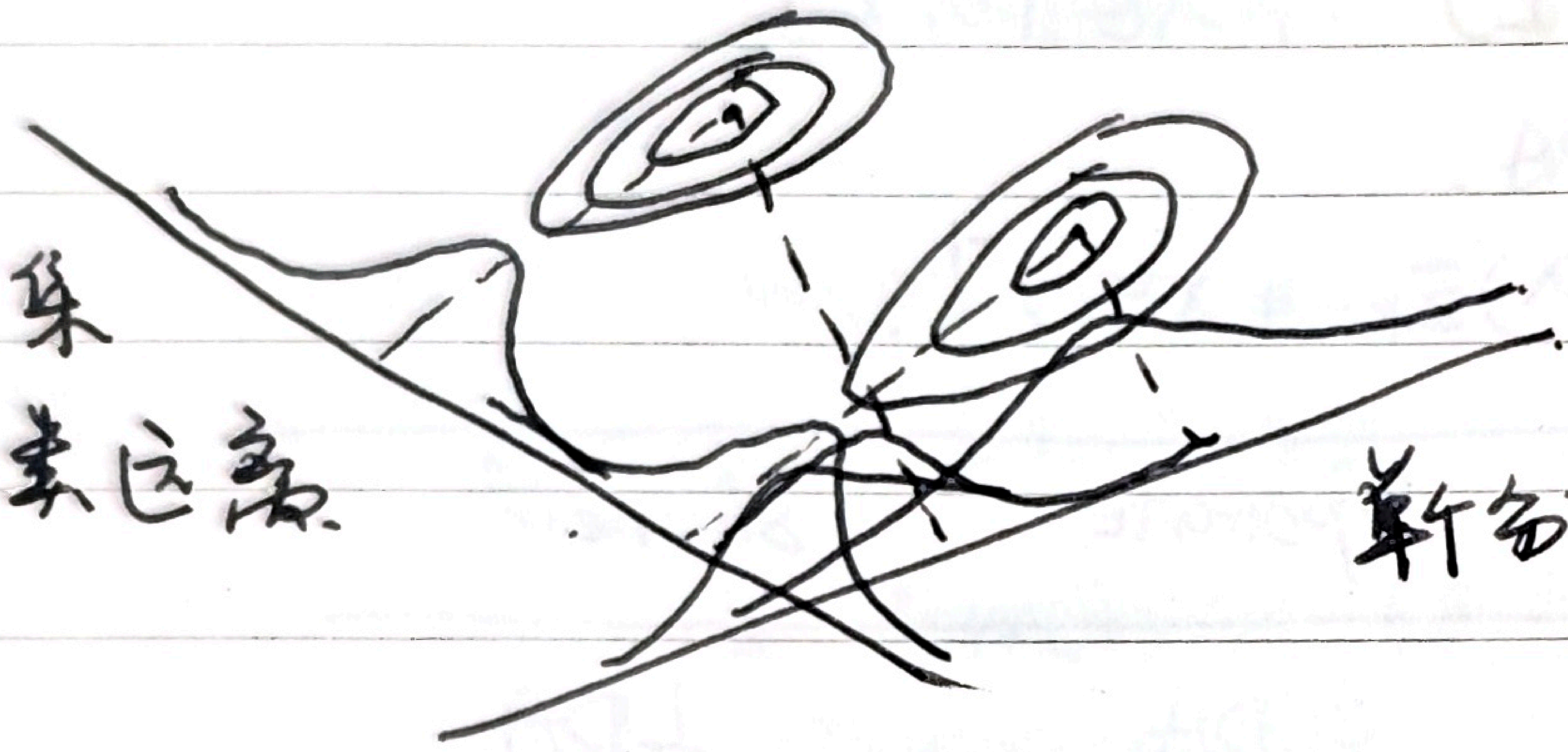
Defination: Fisher's Radio.

$$R = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \qquad \frac{\text{均值之差的平方}}{\text{方差之和}}$$

$$\sim \frac{(w^T(\mu_1 - \mu_2))^2}{w^T \Sigma w}$$

By max R $\mapsto$ $\hat{w} \sim \Sigma^{-1}(\mu_1 - \mu_2)$.

单个分布聚集而两个分类远离



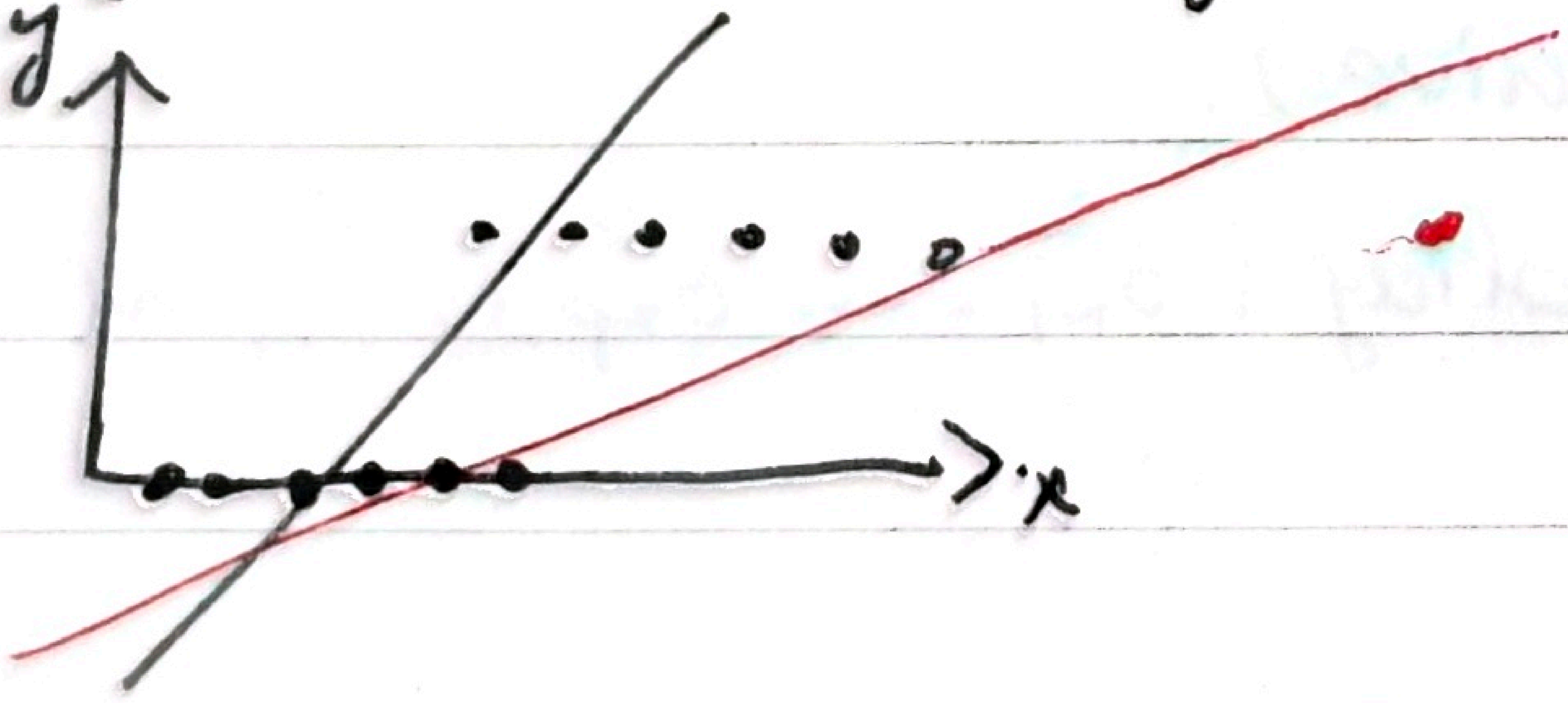
单个分布不聚集而两个分类靠近.

Remark、logistic Regression 和 DA 都可给出后验概率结果!

5. Logistic Regression [Intro to ML Lecture] (概率视角).

A. Classification.

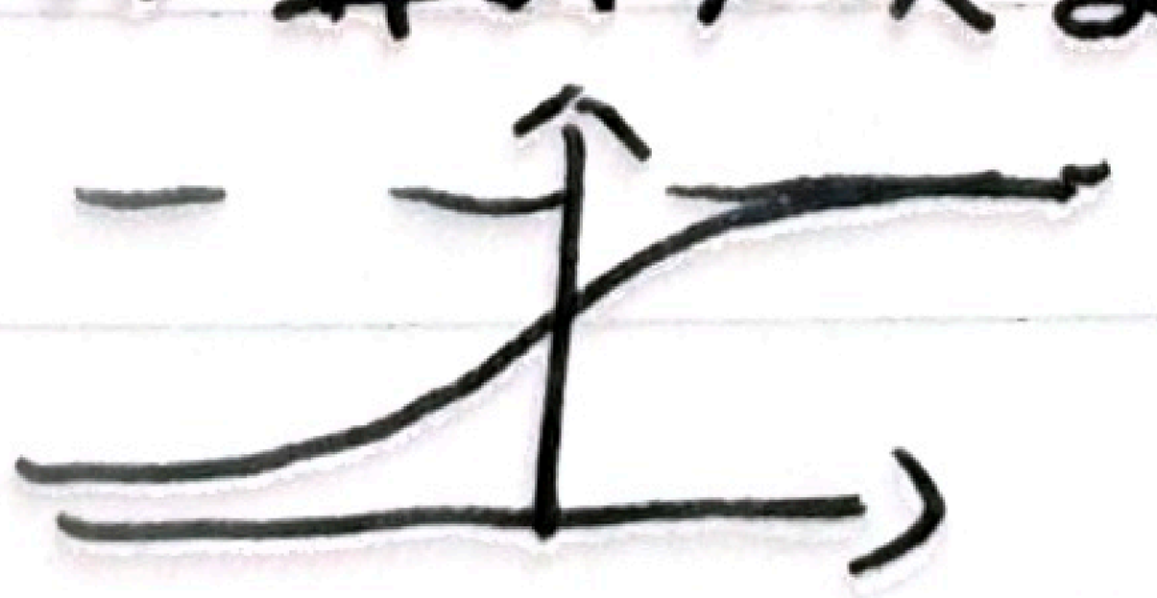
Why not use linear regression?



会因离群点产生奇怪的没有道理的分类情况.

有时,预测概率比直接分类更有意义!

$f(x) = \frac{1}{1+e^{-x}}$ logistic



$\mapsto h(x) = f(\beta^T x) = \frac{1}{1+e^{-\beta^T x}}$.

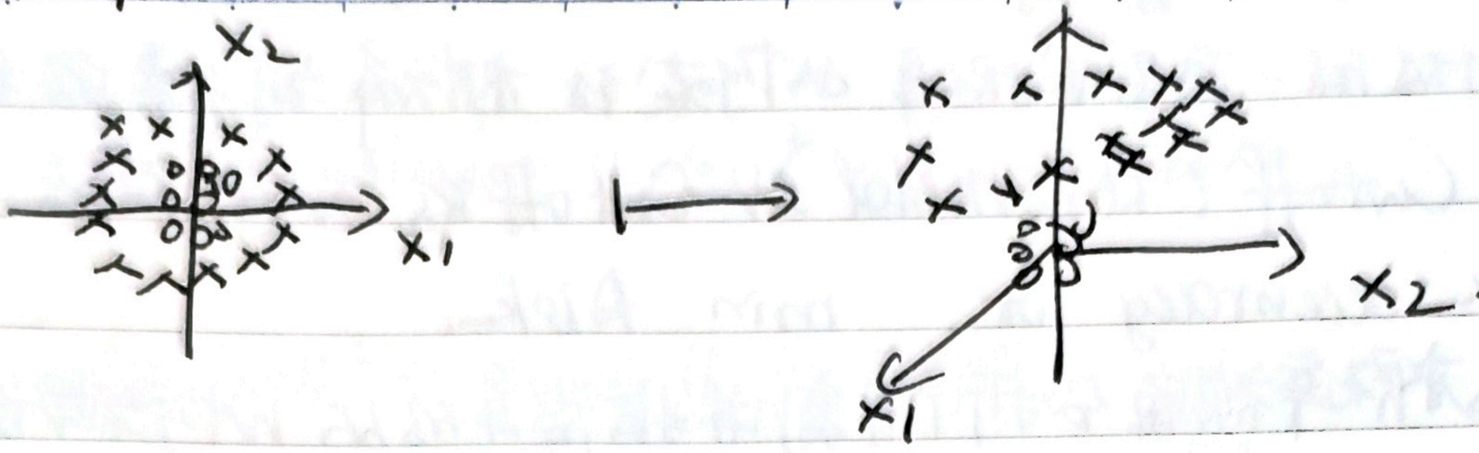
为何 Regression 是 classifier?

[ML] 课程PPT中已有了如何解释, 那就是用 $\beta^T x$ 来当决策边界.

这课程的解释是:线性分类后,再由 logistic 压缩到一个概率.

即为伯努利分布描述概率,来做出参数估计.

启发动画：



通过核函数特征映射使得线性可分！

B. Loss function and model.

Let $\hat{y} = P(y=1) = \frac{1}{1-e^{\beta^T x}} = g(\beta^T x) = h(x)$.

为何不用 MSE 当 loss？ $L = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$.

$\hat{y}_i = 0.pp$ 和 $\hat{y}_i = 0.ppp$ 在 $y_i = 0$ 时有相似的 loss，无法很好的衡量误差.

使用 MLE.

$$L = -\log \text{Likelihood}$$
$$= -\sum_{i|y_i=1} \log h(x_i) - \sum_{i|y_i=0} \log(1-h(x_i))$$
$$= -\sum_i \left[y_i \log h(x_i) + (1-y_i)\log(1-h(x_i)) \right].$$

Generalized linear model, L 是 convex 的，但没有闭式解.

C. 优化.

实际时 logistic 的优化基本两种方法，也主流于 GD.

By $g'(x) = g(x)(1-g(x))$ ~~$f(x) = \frac{1}{1+e^{-x}}$~~ $f(x) = \frac{1}{1+e^{-x}}$

$\Rightarrow$ ~~$\nabla_\beta \log h(x)$~~ $= \nabla \log g(\beta^T x) = x(1-h(x))$

$\nabla_\beta \log(1-h(x)) = -h(x)x$.

$$\nabla L = -\nabla \sum_i [y_i(\log h(x_i)) + (1-y_i)(\log h(x_i)+1)]$$
$$\Rightarrow -\sum_i [x_i(y_i - h(x_i))] \triangleq -X^T(y-\hat{y})$$

D. 补充.

① Overfitting, loss $\mapsto 0$, $\hat{\beta} \mapsto \infty$. $\Rightarrow \hat{y} \to 0$ or $\hat{y} \to 1$.

需要一些调优项面试为 N 折交叉验证.

No.

Date

② Probabilities to binary predictions.

我们在做的一直是求概率，如何将概率划为类的判决呢？

需要一个 Cutoff (Threshold)，Cutoff 的选取原则是：

max accuracy or min Risk.

注：这里从概率视角分析问类，并非简单的从 w 的视角，如此选择 Cutoff 可以融合更多先验信息，使决策更科学。

③ Multinomial logistic regression.

By Softmax! $P(y=k) = \frac{e^{\beta_k^T x}}{\sum e^{\beta_k^T x}}$

第二：从概率视角分析 logistic，与 w 的视角相互照映。

logistic 不止是一种分类技术，其实应该是一种 Regression！

6. 一个关于 k-means 的证明.

Define intra-cluster sample scatter

$$I(C_0) = \frac{1}{2}\sum_{x_i \in C_0} d_{ij} = \frac{1}{2}\sum_{i,j} \|x_i - x_j\|_2^2$$

Define Center (mean) $\mu_0 = \frac{1}{N_0}\sum_i x_i$

Lemma $I(C_0) = \frac{1}{2}\sum_{i,j} \|x_i - x_j\|_2^2 = N_0 \sum_{x_i} d_i$

$$= \underbrace{N_0 \sum \|x_i - \mu_0\|_2^2}_{B}.$$

proof. $A = \frac{1}{2}\sum_{i,j} \|x_i - x_j\|_2^2$

$$= \frac{1}{2}\sum_{i,j}\left[\|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j\rangle\right].$$

$$= \frac{1}{2}\sum_i N_0 \|x_i\|_2^2 + \frac{1}{2}\sum_j N_0 \|x_j\|_2^2 - \sum_{i,j}\langle x_i, x_j\rangle$$

$$= N_0 \sum \|x_i\|_2^2 - \sum_{i,j}\langle x_i, x_j\rangle.$$

$$B = N_0 \sum_1 \| x_i - \mu_0 \|_2^2$$

$$= N_0 \sum_1 \left[\| x_i \|_2^2 + \| \mu_0 \|_2^2 - 2 \langle x_i, \mu_0 \rangle \right]$$

$$= N_0 \sum_1 \| x_i \|_2^2 + N_0 \cdot \sum_1 \| \frac{1}{N_0} \sum_j x_j \|_2^2 - 2 \frac{N_0}{N_0} \sum_1 \langle x_i, \sum_j x_j \rangle$$

$$= N_0 \sum_1 \| x_i \|_2^2 + \frac{N_0 \cdot N_0}{N_0^2} \langle \sum_j x_j, \sum_K x_K \rangle - 2 \sum_{i,j} \langle x_i, x_j \rangle$$

$$= N_0 \sum_1 \| x_i \|_2^2 + \sum_{j,K} \langle x_j, x_K \rangle - 2 \sum_{i,j} \langle x_i, x_j \rangle$$

$$= N_0 \sum_1 \| x_i \|_2^2 - \sum_{i,j} \langle x_i, x_j \rangle = A. \qquad \square$$

使用该 Lemma 可以使用 $I(C_0) = N_0 \sum_1 \| x_i - \mu_0 \|_2^2$ 来简便对消 intra-cluster sample scatter!

7. Kernel Density Estimation.

Empirical distribution: $\hat{f}(x) = \frac{1}{N} \sum_1 \mathbb{I}\{X_i = x\}$.

通过把 $\hat{f}(x)$ 极限化的思想，得到 $\hat{f}(x)$ 的连续型式.

$$f(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h}$$

$$= \lim_{h \to 0} \frac{1}{2h} \left[\frac{1}{N} \sum_1 \mathbb{I}\{ x_0 + h \geq X_i \} - \frac{1}{N} \sum_1 \mathbb{I}\{ x_0 - h \geqslant X_i \} \right]$$

$$= \lim_{h \to 0} \frac{1}{2N} \cdot \frac{1}{h} \left[\sum_1 \mathbb{I}\{ x_0 - h \leq X_i \leq x_0 + h \} \right]$$

$$= \lim_{h \to 0} \frac{1}{Nh} \left[\frac{1}{2} \sum_1 \mathbb{I}\left\{ \left| \frac{X_i - x_0}{h} \right| \leq 1 \right\} \right].$$

$$\triangleq \frac{1}{Nh} \sum_1 k\left(\frac{X_i - x_0}{h} \right), \text{ where } k(\cdot) \text{ is kernel function. } h \text{ is bandwidth.}$$

$$MSE_D\{\hat{f}(x)\} = \mathop{E}_D\left[(\hat{f}(x_0) - f(x_0))^2\right].$$

$$= \left[\mathop{E}_D(\hat{f}(x_0) - f(x_0))\right]^2 + \mathop{Var}_D(\hat{f}(x_0) - f(x_0))$$

$$= \underbrace{\left[f(x_0) - \mathop{E}_D\hat{f}(x_0)\right]^2}_{\text{estimator bias}} + \underbrace{\mathop{E}_D\left[\mathop{E}_D\hat{f}(x) - \hat{f}(x)\right]^2}_{\text{estimator variance}}.$$

通过定义 $MISE = \mathop{E}_D\int_{-\infty}^{+\infty}(\hat{f}(x) - f(x))^2 dx$ 为例证明.

$$= \int bias^2\, dx + \int var^2 dx.$$

当 h 变大时, $\int bias^2 dx$ 变大, $\int var^2 dx$ 变小., 即.

- large h : $\hat{f}(x)$ is smoother (low model flexibility)
  low variance, high bias.
- Small h : $\hat{f}(x)$ is less smoother (high model flexibility, spiky).
  high variance, low bias.

Theorem **Consistency** when $N\to\infty$ $h\to 0$, $Nh\to\infty$, exist
$bias\to 0$, $Var\to 0$. $\hat{f}(x) \xrightarrow{P} f(x)$.

Asymptotic Normality when $N\to\infty$, $h\to 0$, $Nh\to\infty$, exist
$\sqrt{Nh}(\hat{f}(x) - f(x)) \xrightarrow{d} N(0, f(x)\int k^2(z)dz)$.

Note 通过 min MISE 的方法, 可以找到最优 kernel: Epanechnikov.
和最优 $h^* = \arg\min_h MISE(\hat{f})$.