1. Population and Sample

Population: The whole set of individuals about which we attempt to draw conclusions.
Sample: A part of the population which is observed. Sample space (所有)
Relationship: **Probability:** Reasoning from Population →Sample. 随机
Inferential Statistics: Reasoning from Sample →Population. 不是 Population 找

2. Probability Basics & Counting

Definitions 检测到值
Experiment: Any action that generates observations. $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$ $A\cup(B\cup C)=(A\cup B)\cup C$ $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$
Sample Space (S): Set of all possible outcomes.
Event: A subset of outcomes in S. $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$
Probability: $P(A)$ = (# of outcomes in A)/(Total # of outcomes in S) $(A\cup B)^c=A^c\cap B^c$ $(A\cap B)^c=A^c\cup B^c$
Counting Techniques 乘法
Product Rule: Experiment 1 (m outcomes) followed by Experiment 2 (n outcomes) → $m\times n$ outcomes. 排列 $P_n^k=C_n^k\times k!$
Permutation (Ordered): Number of ways to arrange k objects from n distinct objects.
Formula: $P_n^k=n!/(n-k)!$ 组合 $=n\times(n-1)\cdots\times(n-r+1)$ $P_6^3=6\times5\times4$
Combination (Unordered): Number of ways to select k objects from n distinct objects.
Formula: $C_n^k=\binom{n}{k}=n!/(k!(n-k)!)$ $C_5^3=\frac{5!}{3!\times2!}=10$

3. Axioms & Rules

Complement: $P(A')=1-P(A)$. 补集
Mutually Exclusive: If $A\cap B=\emptyset$, then $P(A\cap B)=0$.
Addition Rule: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
Three Events:
$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-[P(AB)+P(AC)+P(BC)]+P(ABC)$$

4. Conditional Probability & Bayes 样本压缩到 B发生范围内 $P(A|B)$

Conditional Probability: $P(A\mid B)=P(A\cap B)/P(B)$.
Multiplication Rule: $P(A\cap B)=P(A)P(B\mid A)$.
Independence: A and B are independent if $P(A\cap B)=P(A)P(B)$ or $P(A\mid B)=P(A)$.
Bayes Theorem:
$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{\sum_{k=1}^{n}P(A_k)P(B\mid A_k)}$$ 后验
5. random variable is any characteristic(s) whose value change from one individual to an other
Numerical: **Location:** Mean, Median, Trimmed mean. **Variability:** Variance, Standard deviation.
Graphical: Histogram (Density/Frequency), Pie chart.
Boxplot: Visualizes Median, 1st/3rd Quartiles, and Outliers.

6. Random Variables (R.V.)

6.1 Discrete R.V. 离散型

PMF: $p(x)=P(S\in S: X(s)=x)$ $CDF: F(x)=P(S\in S: X(s)\le x)$
Expectation: $E(X)=\sum xp(x)$. **Variance:** $Var(X)=E(X^2)-[E(X)]^2$ $=E[X-E(X)]^2$
Common Distributions: **Bernoulli:** $p(x)=p^x(1-p)^{1-x}$.
Binomial: $C_n^x p^x(1-p)^{n-x}$. **Poisson:** $p(x)=\frac{\lambda^x}{x!}e^{-\lambda}$

6.2 Continuous R.V. Probability density function (cumulative distribution F)

PDF ($f(x)$): Probability is area under curve. $P(X=x)=0$. **CDF:** $F(x)=\int_{-\infty}^{x}f(t)dt$.

Expectation: $E(X)=\int_{-\infty}^{\infty}xf(x)dx$. iid: $\bar{x}$ unbias. of μ, $\hat{\sigma}^2=\frac{1}{n-1}\sum(X_i-\bar{X})^2$ unbias of σ^2

Common Distributions: $E=\frac{a+b}{2}$
Uniform: $f(x)=1/(b-a)$. $x\in[a,b]$ 指数 **Exponential:** $f(x)=\lambda e^{-\lambda x}(x>0)$.
Normal: $X\sim N(\mu,\sigma^2)$, $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $E(h(X,Y))=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)f(x,y)dxdy$

7. Joint Distribution

边缘
Joint PDF: $P((X,Y)\in A)=\iint_A f(x,y)dydx$. **Marginal PDF:** $f_X(x)=\int f(x,y)dy$.
Independence: $f(x,y)=f_X(x)f_Y(y)$. 协方差 **Covariance:** $Cov(X,Y)=E(XY)-E(X)E(Y)$.
Correlation: $Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ condition pdf: $f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}$

8. Statistics & Sampling Random variable

Random Sample: $X_1,\ldots,X_n$ are i.i.d. (independent and identically distributed).
Sampling Distribution of Mean: If $X_i\sim N(\mu,\sigma^2)$, then $\bar{X}\sim N(\mu,\sigma^2/n)$.
Central Limit Theorem (CLT): For large n, regardless of population distribution:
$X_1\cdots X_n$ are i.i.d $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\to N(0,1)$ $Y=\sum a_iX_i\sim N(\sum a_i\mu_i,\sum a_i^2\sigma_i^2)$
Linear Combination: If X_i are independent Normals, $\sum a_iX_i$ is Normal. 不独立时 var. $E(Y)=\sum a_i\mu_i$ $Var(\sum a_iX_i)=\sum a_i^2\sigma_i^2$ $Var=\ldots+2\sum a_ia_j cov(X_i,Y_j)$
Statistical inference: find truth on the population based on data get from a sample of population

9. Point Estimation

Point Estimate: A single value (statistic) to estimate a parameter.
Properties: $\hat{\mu}=\bar{x}$ **Unbiased:** $E(\hat{\theta})=\theta$. (e.g., $\bar{X}$ for μ, S^2 for σ^2). $E(\hat{\theta})-\theta$: bias
MVUE: Minimum Variance Unbiased Estimator. (e.g., $\bar{X}$ for Normal μ).
Methods: **Method of Moments (MM):** Equate sample moments $(\frac{1}{n}\sum X_i^k)$ to population moments ($E[X^k]$).
Maximum Likelihood (MLE): Find θ that maximizes likelihood $L(\theta)=f(x_1,\ldots x_n;\theta)$. joint pdf $f(X_1,\cdots X_n|\theta)=f(X_1,\cdots X_n;\theta)$

10. Confidence Interval (CI)

Definition: A random interval that contains the true parameter with prob $1-\alpha$.
Interpretation: In repeated sampling, $100(1-\alpha)\%$ of intervals cover μ.
Formulas (for Mean): A confidence interval (CI) is an interval estimate, computed based on certain statistics, that may contain the unknown population parameter with pre-specified probability.
Known σ: $\bar{X}\pm z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$ Unknown σ: $\bar{X}\pm t_{\alpha/2,n-1}(\frac{s}{\sqrt{n}})$
General Approx CI: $\hat{\theta}\pm z_{\alpha/2}\sigma_{\hat{\theta}}$ $\hat{\theta}\sim N(\theta,\sigma^2_{\hat{\theta}})$

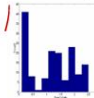
A **hypothesis** is a claim about a characteristic of a probability distribution (population parameter).
A **hypothesis test** uses sample data to choose between two competing claims.
Null hypothesis H_0: assumed true initially. **Alternative hypothesis** H_a: competing claim contradicting H_0.
For $H_0:\theta=\theta_0$, common alternatives are $H_a:\theta>\theta_0$, $H_a:\theta<\theta_0$, or $H_a:\theta\ne\theta_0$.
A **test procedure** is a data-based rule for deciding whether to reject H_0, defined via a **test statistic** (a function of the sample) and a **rejection region** (values for which H_0 is rejected).
Type I error: reject H_0 when H_0 is true; probability $=\alpha$ (significance level). **Type II error:** fail to reject H_0 when H_0 is false; probability $=\beta$. **Power** $=1-\beta$. Typically, increasing α decreases β (increases power).
Goal: minimize β subject to a pre-specified α. P-value: Probability of observing data as extreme as actual, assuming H_0 is true
Recect H_0 if $P<\alpha$

1. What Is Data? 对象

Collection of data **objects** and their **attributes** 属性
Object: record, point, sample, entity, instance.
Attribute: property/characteristic (e.g., age, height). Also called variable, field, feature.

2. Types of Variables

Continuous: length, time, weight. 连续
Nominal (Categorical): race, sex, eye color. 名义
Ordinal: age group, grade, satisfaction rating (ordered). 有序
Interval: temperature, salary range. 区间

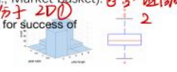


3. Types of Data

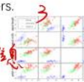
Data Matrix Points in multi-dimensional space.
Text Data: Document = **vector of terms**. Element = term frequency (counts).
Transaction Data: Each record = a set of items (e.g., Market basket). 日常一组商品
Graph Data: Social networks, molecular structures. 分子 2D

4. Data Quality

Quality of data is a very important factor for success of subsequent analysis
Issues affecting analysis:

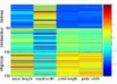


- **Noise:** Perturbation of values.
- **Outliers:** Observations considerably different from others.
- **Missing Values:**
  - Reasons: Not collected or not applicable. to all cases
  - Handling: Eliminate, Impute, or Incorporate partial info. 信息
- **Sampling Bias:** Sample ≠ Population.
  - Causes: Convenience sampling, Class imbalance.



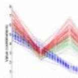
5. Data Exploration (EDA)

- **Goal:** Preliminary understanding. Select right tools.
- **Techniques:** created by John Turkey
  1. **Summary Statistics.**
  2. **Visualization.**
- (Clustering/Anomaly detection are also exploratory).



6. Summary Statistics

- **Frequency/Mode.**
- **Location:** Mean, Median, Trimmed mean, Percentile.
- **Spread:** Range, Variance, Standard deviation. 离散
- **Skewness.**



7. Visualization Techniques

- 1 **Histogram:** Distribution of a variable. (2D Histogram for joint distribution).
- 2 **Boxplot:** Median, quartiles, outliers. Good for comparing distributions.
- 3 **Scatter Plot:** Relationship between 2 variables.
- 4 **Matrix Plot:** Visualizes data/similarity matrix using colors. Variables often normalized.
- 5 **Parallel Coordinates:**
  - Parallel axes for variables.
  - Object = **Line** connecting axes.
  - Variable ordering matters.
- 6 **Star Plots:** Axes radiate from center. Object = Polygon.
- 7 **Chernoff Faces:** Variables mapped to facial features. Object = Face.

1. Terminology

Statistics ↔Machine Learning (Non linear) dimension reduction
Classification/Regression ↔Supervised Learning manifold learning
Clustering ↔Unsupervised Learning X_i: feature vectors, inputs
Covariates ↔Features $X\in R^p$, qualitative features are coded
Sample ↔Training set /testing set y_i: output, response, dependent variable
Statistical Model ↔Learner scalar/real vector
Misclassification/Prediction error → Generalization error random error

2. Models & Assumptions

Assumption 19: $Y=f(X)+\epsilon$, where $E(\epsilon)=0$ and is independent of X.

Model Types 20
Parametric (Strong assumption on f): Linear/Polynomial regression, GLM, Fisher's discriminant, Logistic regression, Deep learning.
Nonparametric (Flexible form): Local smoothing, Smoothing splines, Trees/Random Forest/Boosting, SVM.

3. Prediction vs Inference

$E(\hat{f}(X)-f(X)-\epsilon)^2=E(\hat{f}(X)-f(X))^2+var(\epsilon)$
Prediction: Minimize $E(\hat{Y}-Y)^2$ = Reducible Error + Irreducible Error $Var(\epsilon)$.
Inference: Understand relationship between X and Y. fact
Trade-off: Linear models (High Inference, Low Prediction accuracy) vs Non-linear models (Low Inference, High Prediction). decision F: $\hat{\phi}(x)=\arg\max_k \hat{f}_k(x)$

4. Classification Models Detail

Task: Estimate $P(Y=k\mid X)$. ; $k=1,\cdots K$ error: $P(Y\neq\hat{\phi}(X))$ $=E(I(Y\neq\hat{\phi}(X)))$ I(真):1 假:0

Model 1: Linear Regression

Method: Fit Y(0 or 1) using LSE $\hat{\beta}=(X^TX)^{-1}X^Ty$. $Y=\beta_0+\sum X_j\beta_j=X\beta$
Decision: $\hat{\phi}(X)=I(X^T\hat{\beta}>0.5)$. $RSS(\beta)=\sum(y_i-x_i^T\beta)^2=(y-X\beta)^T(y-X\beta)$
(Added) Trivial classifier example: I(Feature $>C$)23.

Model 2: K-Nearest Neighbors (KNN)

N_k: neighbour hood (包含k点)
Method: $\hat{y}(X)=\frac{1}{k}\sum_{i\in N_k}y_i$. $=\frac{1}{k}\sum_{i=1}^{n}y_iI(X_i\in N_k(X))$
Decision: Majority vote. $\hat{\phi}(X)=I(\hat{y}(X)>0.5)$
Effective Parameters: n/k.
$k=1$: 0 training error (Overfitting), rough boundary.
Large k: Smoother boundary. K选奇数防平票

5. Model Assessment

5.1 Regression (MSE) mean square error.

Measure: $MSE(f)=E(Y-f(X))^2$.
Optimal Solution: The Conditional Expectation $f^*(X)=E(Y\mid X)$ minimizes MSE.
(X,Y) distribution is unknow, MSE from sample data.

5.2 Classification (MCE)

$(x_i,y_i)\in$ train set
Measure: Misclassification Error $E(I(Y\neq f(X)))$. $\frac{1}{n}\sum(y_i-\hat{f}(x_i))^2$
Optimal Solution: Bayes Rule $f^*(X)=\text{argmax}_kP(Y=k\mid X)$. 估计 MSE(f)

5.3 Training vs Test → not seen during training

Training error under-estimates true error. (X_{0i},y_{0i}): test
We need Test error (from independent data). $\frac{1}{m}\sum(y_{0i}-\hat{f}(X_{0i}))^2$ test error → MSE than train.
$var(\hat{f}(X))=E(\hat{f}(X)-E[\hat{f}(X)])^2$ ↑过拟合

6. Bias-Variance Decomposition

Test MSE decomposition: $Bias(\hat{f}(X))=E(\hat{f}(X))-f(X)$ 欠拟合
$E(Y-\hat{f}(X))^2=E(\hat{f}(X)-f(X))^2+var(\epsilon)=(Bias(\hat{f}(X)))^2+var(\hat{f}(X))+var(\epsilon)$
Bias: Error from simplifying assumptions (High in simple models).
Variance: Sensitivity to training set fluctuations (High in complex models).
Trade-off: Complexity ↑⇒Bias ↓, Variance ↑. misclassification error (MCE)

7. Cross Validation (CV) & Estimation

$MCE(f)=E(I(Y\neq f(X)))$
Alternative to Test Set: (minimizer of MCE) $f^*(X)=\arg\max_k P(Y=k|X)$ (Bayes rule) $\frac{1}{n}\sum I(y_i\neq f(x_i))$ train error.
调 test error
Adjustment29: Adjust training error using AIC, BIC, Covariance penalty.
Resampling Methods: (没有大量数据在现实中时)

Validation Set
Split data randomly. into training and validation parts
Cons: High variance in result, overestimates error (less training data).
Ad: simple idea, easy to implement. 实现

Leave-One-Out CV (LOOCV) MSE = averaged MSE
Train on $n-1$, Test on 1. Repeat n times.
Pros: Low Bias. less variable MSE 比起普通的 $\frac{k}{N}$ 做为验证集
Cons: Computational cost. 训练K次模型
$E_k(\hat{f})=\sum_{i\in A_k}L(y_i,\hat{f}^{-k}(x_i))^2$ K折

K-Fold CV
Split into K folds. $CV(\hat{f})=\frac{1}{n}\sum_{k=1}^{k}E_k(\hat{f})$ $(A_1\cdots A_k)$
CV = Average of E_k.

Comparison
Bias: LOOCV < K-Fold.
Variance: LOOCV > K-Fold. (when k<n)
Choice: $K=5$ or 10 provides the best compromise.

1. 简单线性回归（Simple Linear Regression）

基本模型
定义：$f(x)=E(Y\mid X=x)$。线性模型假设 $f(x)$近似为线性函数。公式：
$y_i=\beta_0+\beta_1x_i+\epsilon_i$ independent V dependent variable
x_i：预测变量（Predictor/Input/Feature）。y_i：响应变量（Response/Output/Outcome）。
最小二乘拟合（Least Squares Fitting） 2
目标：最小化残差平方和 $\arg\min\sum(y_i-\beta_0-\beta_1x_i)^2$。$=(\hat{\beta}_0,\hat{\beta}_1)$
解（Solution） $\hat{\beta}_1=\frac{\sum_{(i=1)}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{(i=1)}^{n}(x_i-\bar{x})^2}$ $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ (过数据中点)
点估计与假设 $e_i=y_i-\hat{y}_i$ (residuals)
假设：误差项 ϵ_i独立同分布于 $N(0,\sigma^2)$。
$\hat{\beta}_0$和 $\hat{\beta}_1$是 y_i的函数，因此也是随机变量。分布性质：
最小二乘 $\hat{\beta}_1\sim N(\beta_1,\frac{\sigma^2}{\sum(x_i-\bar{x})^2})$ $\hat{\beta}_0\sim N(\beta_0,(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2})\sigma^2)$ $b_1=\sum C_iy_i$
BLUE 性质：$\hat{\beta}_1$是所有无偏线性估计量中方差最小的（Best Linear Unbiased Estimator）。

2. 推断：置信区间与检验（Inference: CIs & Tests） BLUE ⊆ MVUE

方差估计 σ^2的无偏估计（MSE）: $\hat{\sigma}^2=\frac{\sum(y_i-\hat{y}_i)^2}{n-2}$
分布：$(n-2)\hat{\sigma}^2/\sigma^2\sim\chi^2(n-2)$。
系数的置信区间 (CI) 6
由于 $\hat{\beta}$和 $\hat{\sigma}^2$独立（Cochran 定理），构建 t 统计量。
β_1的 $100(1-\alpha)\%$CI: $\hat{\beta}_1\pm t(\frac{\alpha}{2},n-2)S(\hat{\beta}_1)$
其中标准误 $s(\hat{\beta}_1)$ $(\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2})^{\frac{1}{2}}$ $s(\hat{\beta}_0)$ $(\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2\bar{x}^2}{\sum(x_i-\bar{x})^2})^{\frac{1}{2}}$ CI 同上
假设检验 H_0
检验 $\beta_1=0$(是否存在线性关系):
统计量：$t_1^*=\hat{\beta}_1/s(\hat{\beta}_1)\sim t_{n-2}$
决策：如果 $|t_1^*|>t(\alpha/2,n-2)$则拒绝 H_0。t_0 同理
拟合值的推断
拟合直线 $\hat{f}(x)$在某点 x处的标准误 $s(\hat{f}(x))$: $S(\hat{f}(x))=(\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})^{\frac{1}{2}}$
这构成了拟合线的置信带（Confidence band）。

3. 多元线性回归（Multiple Linear Regression, MLR）

矩阵表示
模型：$y_i=\beta_0+\sum\beta_jx_{ij}+\epsilon_i$。 矩阵形式：$y=X\beta+\epsilon$。
X: $n\times(p+1)$矩阵（第一列全为 1）。 β: $(p+1)$维参数向量。
最小二乘解
目标：最小化 $(y-X\beta)^T(y-X\beta)$。
估计量：
$$\hat{\beta}=(X^TX)^{-1}X^Ty$$
性质:无偏：$E(\hat{\beta})=\beta$ 协方差矩阵：$Cov(\hat{\beta})=\sigma^2(X^TX)^{-1}$
拟合值与 Hat Matrix: $\hat{y}=X\hat{\beta}=X(X^TX)^{-1}X^Ty=Hy$ $E(\hat{y})=X\beta$, $cov(\hat{y})=\sigma^2H$
H被称为 Hat Matrix (帽子矩阵)，将 y投影到 X的列空间。
残差性质
残差向量：$e=y-\hat{y}=(I-H)y$。$E(e)=0$
期望为 0，协方差 $Cov(e)=\sigma^2(I-H)$。 $E(SSE)=\sigma^2(n-p-1)$
残差平方和期望：$E(e^Te)=\sigma^2(n-p-1)$。 $E(SSR)=p\sigma^2+\beta^TX^T(I-\frac{J}{n})X\beta$ $=E(SSTO)$
σ^2的估计：$\hat{\sigma}^2=MSE=e^Te/(n-p-1)$。

4. 方差分析与模型评估（ANOVA & Evaluation） ANOVA 分解

$$SSTO=SSE+SSR$$
SSTO (Total)：总变异 $y^T(I-\frac{1}{n}J)y$。 SSE (Error)：误差平方和 $e^Te=y^Ty-\hat{\beta}^TX^Ty$
SSR (Regression)：回归平方和 $y^T(H-\frac{1}{n}J)y=y^Ty-\hat{\beta}^TX^Ty$
R^2与 调整 R^2 13
R^2: $SSR/SSTO=1-SSE/SSTO$。
随着变量增加，R^2 总是增加，不适合模型比较
Adjusted $R^2(R_a^2)$: 考虑了变量个数 p的惩罚。$1-\frac{\frac{SSE}{n-p-1}}{\frac{SSTO}{n-1}}=1-\frac{n-1}{n-p-1}\frac{SSE}{SSTO}$
F 检验（Global Test）
假设：$H_0:\beta_1=\ldots=\beta_p=0$(所有斜率为 0)。 统计量：$F^*=\frac{MSR}{MSE}=\frac{SSR/P}{SSE/(n-P-1)}$
决策：如果 $F^*>F(\alpha;p,n-p-1)$则拒绝 H_0。

5. MLR 的系数检验（Tests for Coefficients） 单个系数的 t 检验

用于检验某个特定变量 X_k是否显著 $(H_0:\beta_k=0)$。covariance $=\sigma^2(X^TX)^{-1}$
$\hat{\beta}$的估计协方差矩阵：$s^2(\hat{\beta})=MSE(X^TX)^{-1}$
统计量：（公式）$\frac{\hat{\beta}_k-\beta_k}{S(\hat{\beta}_k)}\sim t(n-P-1)$ $t^*=\hat{\beta}_k/S(\hat{\beta}_k)$
置信区间：$\hat{\beta}_k\pm t(\alpha/2,n-p-1)s(\hat{\beta}_k)$。 $|t^*|>t(\frac{\alpha}{2},n-P-1)$，拒 $H_0(\beta=0)$ 单变量 T，多变量 F

6. 模型诊断（Model Diagnostics） 潜在问题

1.非线性（Nonlinearity）。2.遗漏重要变量。3.异方差性（Non-constant variance）。4.误差不独立（Non-independence）。5.误差非正态（Non-normality）。6.离群点(Outliers)。7.多重共线性（Multicollinearity）。
残差图（Residual Plots）17 semi-studentized residuals
使用残差 e_i或半学生化残差 $e_i^*=e_i/\sqrt{MSE}$。$=e_i-\bar{e}/\sqrt{MSE}$
检测非线性：绘制 $e_i$vs $\hat{y}_i$。如果有曲线模式，说明存在非线性（线性假设不成立）
检测遗漏变量：绘制 $e_i$vs 其他预测变量 。Heteroscedasticity
检测异方差：绘制 $|e_i|$vs $\hat{y}_i$。如果呈现漏斗形（funnel shape），说明方差不恒定 。
检测依赖性：绘制 $e_i$vs 时间/空间序列。如果相邻点相关，说明误差不独立 。
检测非正态性：QQ Plot (Quantile-Quantile plot)。如果点偏离直线，说明非正态 。

7. 离群点与高杠杆点（Outliers & High Leverage）

离群点（Outlying Y） $e_i=y_i-\hat{y}_i$ $e_i^*=\frac{e_i}{\sqrt{MSE}}$ (semi-studentized residuals)
Y值远离预测值的点。检测指标：
学生化残差：$r_i=e_i/(s(e_i))$。 $(=y_i-\hat{y}_{i(-i)})$
删除残差（Deleted residual）d_i：不使用第 i个点拟合模型时的预测误差。$d_i=\frac{e_i}{1-h_{ii}}$
学生化删除残差（Studentized deleted residual） $t_i=e_i\left(\frac{n-p-2}{SSE(1-h_{ii})-e_i^2}\right)^{\frac{1}{2}}\sim t_{n-p-2}$
检验：比较 $|t_i|$与 Bonferroni 修正后的 t 值 24。
高杠杆点（Outlying X）25252525
X值异常的点。
指标：Hat Matrix 的对角元素 h_{ii}(Leverage)。$=tr(H)=P+1$
阈值：如果 $h_{ii}>2(p+1)/n$，通常被视为高杠杆点。

8. 多重共线性（Multicollinearity）

定义与后果 26
定义：预测变量之间存在高度相关性。
后果： 1.系数估计的方差变大（不稳定）。2.删除一个变量可能导致系数符号改变。显著性可能被掩盖。3.回归中预测变量的边际显著依赖其他预测变量
检测（VIF）27
方差膨胀因子（VIF）: $(VIF)_j=(1-R_j^2)^{-1}$ 第j个变量对其余 P-1 回归
其中 R_j^2是 X_j对其他所有预测变量回归的 R^2。
阈值：最大 VIF > 10 或 平均 VIF 远大于 1 表示严重共线性。

9. 变换与偏差-方差权衡（Transformation & Bias-Variance Tradeoff）

变量变换（Box-Cox）28
目的：线性化回归函数、稳定方差、正态化误差。
使用 y^λ变换，通过最大似然估计选择最佳 λ。$L(\lambda;\beta_0,\beta_1,\sigma^2)=\frac{1}{(2\pi\sigma)^{n/2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^\lambda-\beta_0-\beta_1^Tx_i)^2)$
偏差-方差权衡（Bias-Variance Tradeoff）29
MSE 分解：$MSE(\hat{f}(x))=Var(\hat{f}(x))+[Bias(\hat{f}(x))]^2$
LSE估计：无偏，但方差可能较大。(线性假设正确)
正则化（Regularization）：如 Ridge/Lasso，通过引入少量偏差来大幅降低方差，从而降低总 MSE。

10. 定性变量与交互项（Qualitative & Interactions） 虚拟变量（Dummy Variables）

对于 2 个水平的定性变量，定义 $X_2=1$(Level 1) 或 0(Level 2)。
模型：$E(Y\mid X)=\beta_0+\beta_1X_1+\beta_2X_2$。
解释：β_2是两个水平之间平均响应的截距差（平行线）。
交互效应（Interaction Effects） 模型：加入乘积项 $\beta_3X_1X_2$。
$E(Y\mid X)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$。
解释：允许不同水平有不同的斜率（非平行线）。
β_3是斜率的差异。多水平编码 5 个水平需要 4 个虚拟变量。
编码方式: 0/1 编码 或 1/-1 编码

1. Sparse Regression & Motivation Basic Setup 1

Given training set (x_i,y_i), assume the linear model:
稀疏性 P_0个有用 $y_i=\beta_0+\sum\beta_jx_{ij}+\epsilon_i$ $A^*=\{P_1,P_2\cdots P_0\}$
Sparsity: It is assumed that $p_0\ll p$, where p_0 is the number of informative predictors.
Goal: Correctly detect the set of informative predictors A^* from the total set of variables.
Why Do We Care? 2 掩盖显著性 膨胀
Multicollinearity: Can mask significance and inflate variance.
Prediction Accuracy: When p is large, accuracy deteriorates due to overfitting ("curse of dimensionality"). 恶化 解释
Interpretability: Models are unnecessarily complicated if irrelevant variables are included.
Popular Techniques 3
Best subset selection. (various information criteria, cross validation
Sequential selection (Forward/Backward). 序贯变量选择
Shrinkage methods (Lasso, Ridge). 充分降维
Dimension reduction (PCA). Principal component analysis, sufficient dimension reduction

2. Subset Selection Algorithms

2.1 Best Subset Selection 4
Let M_0 be the null model (no predictors).
For $k=1,2,\ldots,p$: → 特征数
Fit all $\binom{p}{k}$ models with exactly k predictors. 对每个k选出 M_k，k特征下最好的
Pick the best one (smallest RSS or largest R^2), call it M_k.
Select the single best model from $M_0,\ldots,M_p$ using cross-validation, C_p, AIC, or BIC.
Note: Computationally expensive for large p. ←不能只比RSS.残·平和，变量↑,RSS↓
2.2 Model Selection Criteria 5
Used to select the final model size (d predictors). $SSE\to SSE_d$
Mallow's Cp:
$$C_p=(1/n)*(SSE_d+2d\hat{\sigma}^2)$$ n:样本量
AIC (Akaike Information Criterion):
$$AIC=(1/(n\hat{\sigma}^2))*(SSE_d+2d\hat{\sigma}^2)$$
BIC (Bayesian Information Criterion):
$$BIC=(1/(n\hat{\sigma}^2))*(SSE_d+\log(n)d\hat{\sigma}^2)$$
Note: BIC penalizes complexity more heavily than AIC (since $\log(n)>2$ for $n>7$).
Other: Adjusted R^2, Cross-Validation (estimates test error directly).
2.3 Sequential Variable Selection 6 BIC→变量少,更简单的模型 AIC→变量多,更复杂
Forward Stepwise Selection:
Start with null model M_0.
For $k=0,\ldots,p-1$: Consider all $p-k$ models that add one predictor to M_k. Choose the best (smallest RSS) as M_{k+1}. 加最好的特征
Select best model among $M_0\ldots M_p$ using criteria (AIC/BIC). 踢最没用的
Backward Selection: Starts with full model M_p and iteratively deletes predictors.
Stagewise Selection: Mixes forward addition and backward deletion in each step. 加新 踢出 不显著的p
Remarks 7: Computationally more efficient than Best Subset. selection
Greedy Algorithm: No guarantee of finding the globally best model, but performs well in practice. Forward/backward (centralization 后)

3. Shrinkage Methods (Regularization) $RSS=\sum(y_i-x_i^T\beta)^2+\lambda J(\beta)$

General Formulation 8 $(\hat{\beta}_0,\hat{\beta})=argmin\sum(y_i-\beta_0-x_i^T\beta)^2+\lambda J(\beta)$ 调节参数
λ: Tuning parameter controlling the trade-off between fitting and shrinkage.

The objective function for a linear SVM is: $\min_{\beta_0,\beta}\frac{1}{2}\|\beta\|^2$
Subject to: $y_i(\beta^Tx_i+\beta_0)\ge1$
Nonlinear SVM (Soft Margin SVM):
For nonlinear separability, we introduce slack variables ξ_i to allow some data points to violate the margin, which results in the following optimization function:
$\min_{\beta_0,\beta,\xi}\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^{n}\xi_i$ Subject to: $y_i(\beta^Tx_i+\beta_0)\ge1-\xi_i,\ \xi_i\ge0\ \forall i$

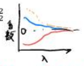
Through this optimization problem, SVM seeks the best hyperplane that classifies the data with maximum margin while allowing some margin violations through the slack variables.
调参 (C)：C 控制松弛变量的大小，即容忍分类错误的程度。较大的 C 会使得模型对错误点不宽容，容易过拟合；较小的 C 会使得模型更宽容，可能有更好的泛化能力。
C controls the magnitude of the slack variables, i.e., the degree of tolerance for classification errors. A larger C makes the model less tolerant of misclassified points, which tends to cause overfitting; a smaller C makes the model more tolerant, potentially yielding better generalization ability.

3.1 Ridge Regression 9

Penalty: L_2-norm penalty $(J(\beta) = \|\beta\|_2^2 = \sum\beta_j^2)$.

Objective: $\hat{\beta}^{ridge} = argmin(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_2^2$

$\Rightarrow argmin(y-X\beta)^T(y-X\beta)$ s.t. $\|\beta\|^2 \le s$



Solution (Closed-form exists):

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

Properties: $\lambda\|\beta\|_2^2$ is a shrinkage penalty shrinks estimates of $\beta \to 0$

Shrinks coefficients towards zero but not exactly to zero (does not perform variable selection).

If $\lambda = 0$, result is OLS (Least Squares). If $\lambda \to \infty$, $\beta \to 0$.

Bias-Variance Trade-off: Ridge is a biased estimator but may have smaller MSE than OLS by reducing variance.

3.2 Lasso (Least Absolute Shrinkage and Selection Operator) 11

Penalty: L_1-norm penalty $(J(\beta) = \|\beta\|_1 = \sum|\beta_j|)$.

Objective:

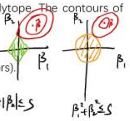
$$\hat{\beta}^{lasso} = argmin(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_1$$

Solution: No explicit closed-form solution (requires Quadratic Programming). 无公式解

Key Property: 会自动为0

Sparse Solution: Some coefficients become exactly zero. Lasso performs continuous variable selection

Geometry: The constraint region $(\sum|\beta_j| \le s)$ is a diamond/polytope. The contours of RSS often touch the "corners" (where coefficients are zero) 13.



3.3 Comparison: Ridge vs. Lasso 14

Both introduce bias to reduce variance. 牺牲

Lasso produces simpler, interpretable models (subset of predictors).

Ridge keeps all variables. (不知哪个更好)

Prediction accuracy depends on the true underlying model.

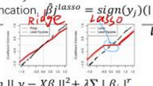
3.4 Special Simple Case ($n = p, X = I$) 15

If the design matrix is orthogonal ($X = I_p$): 正交

OLS Estimate: $\hat{\beta}_j^{ols} = y_j$

Ridge Estimate: Scaling. $\hat{\beta}_j^{ridge} = y_j/(1 + \lambda)$

Lasso Estimate: Soft-thresholding / Truncation. $\hat{\beta}_j^{lasso} = sign(y_j)(|y_j| - \lambda/2)_+$



4. Extensions of Shrinkage Methods

4.1 Bridge Estimators 16

General penalty $L_r(\beta) = \sum|\beta_j|^r$.

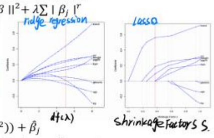
$$\hat{\beta}^{bridge} = argmin\|y - X\beta\|^2 + \lambda\sum|\beta_j|^r$$

Cases: $r = 0$: Hard thresholding (Subset selection). (AIC, BIC)

$r = 1$: Lasso. $L_1(\beta) = \sum_{j=1}^{p}|\beta_j|$

$r = 2$: Ridge. $L_2(\beta) = \sum_{j=1}^{p}\beta_j^2$ (R-regression)

$r = \infty$: Maximum absolute value. $L_\infty(\beta) = \max_j \beta_j$



4.2 Nonnegative Garrote 17 $L_0(\beta) = \sum_{j=1}^{p} I(\beta_j \neq 0)$

Starts with OLS estimates and shrinks them.

Formula: $\min \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_j c_j\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p} c_j$, $c_j \ge 0$, $\hat{\beta}_j^{ng} = \hat{c}_j\hat{\beta}_j$ $\Rightarrow$

$$\hat{\beta}_j^{ng} = (1 - \lambda/(2\hat{\beta}j^2))_+\hat{\beta}_j$$

Property: Almost unbiased for large coefficients, shrinks small ones to zero.

4.3 Other Variants 18

Group Lasso: If variables are partitioned into groups, it includes/excludes whole groups (uses L_2 norm of the group vector). $\min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda\sum_{j=1}^{p}\|\vec{\beta}_j\|_2$ β_j: 系数向量 j group

Elastic Net: Combination of L_1 and L_2 penalties. Elastic: $+\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$

Fused Lasso: Penalizes differences between adjacent coefficients $\min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda\sum_{j=2}^{p}\|\beta_j - \beta_{j-1}\|_1$

5. Dimension Reduction Methods

5.1 Principal Component Analysis (PCA)

Definition 19:

Find linear combinations $U_j = a_j^TX$ (Principal Components) such that:

Variance is maximized: $Var(U_j)$ is as large as possible.

Uncorrelated: $Cov(U_j, U_l) = 0$. PC_1: 数据分布最大的方向, 投影, 方差 max

Normalized: $\|a_j\| = 1$. PC_2: 与 PC_1 ⊥, 剩余方向中变异最大的

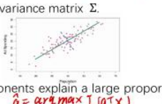
Calculation 20: 特征值分解

Based on Eigen-decomposition of the covariance matrix Σ.

Eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$.

PC loadings a_j are the eigenvectors e_j.

Variance of j-th PC is λ_j.



Dimension Reduction:

Choose $k \ll p$ such that the first k components explain a large proportion α of the total variance. $\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \ge \alpha$

5.2 Projection Pursuit (PP) 22

$\hat{a} = \arg\max_a I(a^TX)$, I: $I_1(z) = |k_m(z)|/k_2(z)^{\frac{m}{2}}$, $I_2(z) = k_3^2(z) + k_4^2(z)/4$

$k_m(z)$: m-th order cummulant of z

Motivation: PCA finds max variance (often Gaussian/Normal). PP seeks directions that are non-normal. 找最有结构方向的

Indices: Measures of non-normality (e.g., using cumulants $k_m(z)$). kurtosis 峰度

5.3 Principal Component Regression (PCR)

$z_j = Xq_j$; $\hat{\gamma}$ is the correlation coefficient of y on z

Algorithm 23: $z = q_1x_1 + q_2x_2 \cdots$

Compute Principal Components $Z_1, \ldots, Z_J$ from X.

Regress response y on these derived components Z_j.

$$\hat{y}^{pcr} = \bar{y} + \sum_{j=1}^{J}\hat{\gamma}_j z_j$$

Since Z_j are orthogonal (正交), this is a sum of univariate regressions. 单独 → Univariate

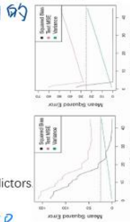
Remarks 24:

Unsupervised: PCA is performed on X without knowing y.

No Variable Selection: Each PC is a linear combination of all original predictors.

Standardization: Recommended to standardize predictors first.

Works well when the first few PCs capture the variation relevant to y.

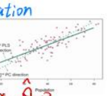


PCR: number of PC chosen by cross-validation

5.4 Partial Least Squares (PLS)

Difference from PCR: 最小偏二乘法

PLS is Supervised. It uses both y and X to construct the directions 25.



Algorithm Concept 26: PLS:

Calculate univariate regression coefficient of y on each x_j.

Use these coefficients to weight inputs and construct the first direction Z

Regress y on Z_1. $y \sim Z_1 \Rightarrow \hat{\beta}_1$, $R_1 = Y - \hat{\beta}_1 Z_1$ (Y的残差), $x_j^* = x_j - \hat{\theta}_j Z_1$

Orthogonalize X with respect to Z_1 (remove effect of Z_1) and repeat to find subsequent directions. 正交化. θ_j: x_j 对 Z_1 回归的系数

Properties 27: 一直到 Z_j

If $J = p$ directions are used, it equals the OLS solution.

Uses $J < p$ for dimension reduction.

The inputs are weighted by the strength of their univariate effect on y.

1. Polynomial Regression 线性回归

Concept Extends linear regression by replacing the linear function with a polynomial. It is a special case of multiple linear regression.

Formula:

$$y_i = \beta_0 + \beta_1x_i + \beta_2x_i^2 + \ldots + \beta_dx_i^d + \epsilon_i$$

Practice: Degree d is usually 2, 3, or 4.

Centered predictors $(\tilde{x}_i = x_i - \bar{x})$ are used to reduce correlation between terms.

Polynomial Logistic Regression

Used for binary responses.

Formula:

$$logit(Pr(y_i > C \mid x_i)) = \beta_0 + \beta_1x_i + \ldots + \beta_dx_i^d$$

2. Step Functions 19

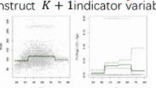
Idea 局部逼近非线性结构 连续 → 二元

Approximates nonlinear structure locally by converting a continuous variable into an ordered binary variable.

Uses K ~~cutpoints~~ $c_1, \ldots, c_K$ to define intervals.

Model breaks points

Construct $K + 1$ indicator variables:



$$C_0(X) = I(X < c_1)$$
$$C_1(X) = I(c_1 \le X < c_2)$$
$$\ldots$$
$$C_K(X) = I(X \ge c_K)$$

3个类只能2组哑变量

Formula:

$$y_i = \beta_0 + \beta_1C_1(x_i) + \ldots + \beta_KC_K(x_i) + \epsilon_i$$

This results in a piecewise constant function. 水平线 β_0 (if $x < c_1$)

3. Interaction Regression Model 交互效应

Used when the effect of one predictor depends on the value of another predictor.

Formula: $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_{12}x_{i1}x_{i2} + \epsilon_i$

The mean response change per unit change in each variable now depends on other variables. 取决于其他变量

regression $b_k(x) = x^k$; step $b_k(x) = I(C_k \le x < C_{k+1})$

(改进阶梯) 分段多项式

4. Piecewise Polynomials & Splines 4.1 Basics

$y_i = \beta_0 + \beta_1 b_1(x_i) \cdots + \beta_k b_k(x_i) + \epsilon_i$

Piecewise Polynomials: Fit separate low-degree polynomials over different regions defined by knots. 在节点定义的不同区域上拟合多个低次多项式

Issue: Discontinuous (jumps) at knots if unconstrained. ↑ flexible↑

4.2 Constraints 22 平滑

① To fix discontinuity, require the fitted curve to be continuous.

② Furthermore, require continuous 1st and 2nd derivatives are continuous.

4.3 Cubic Spline 三次样条

Definition: Continuous piecewise polynomials with continuous 1st and 2nd order derivatives. 导数

Basis Functions: Uses Truncated Power Basis $h(x, \xi)$.

$y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots \beta_{k+3} b_{k+3}(x_i) + \epsilon_i$

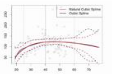
Where $(x - \xi_k)_+^3$ is $(x - \xi_k)^3$ if $x > \xi_k$, else 0. $= \max((x-\xi_k)^3, 0)$

4.4 Natural Cubic Spline

Motivation: Regular splines have high variance at the boundaries.

Definition: A spline require the model to be linear at boundary.

Benefit: More stable estimates at boundaries.



4.5 Determining Knots 百分位

Placement:

Uniform fashion. 均匀放 At percentiles of the predictor (e.g., 25th, 50th, 75th).

Number: Determines complexity (Degrees of Freedom). Chosen via Cross Validation.

5. Smoothing Spline

Objective Function

Find g that minimizes: 自己选节点个数

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\int(g''(t))^2dt$$

Loss: Fits the data. (loss function) Penalty: Penalizes roughness (g'' is curvature). 曲率

Properties Linear in the region outside.

The minimizer $\hat{g}(x)$ is a Natural Cubic Spline. each knots 1阶 2阶 可导

Knots: Every unique observation x_i is a knot.

It is a shrunken version of the natural cubic spline.

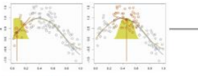
Tuning Parameter λ different $\lambda \to$ different $\hat{g}_\lambda$, all the training observations

Controls bias-variance trade-off.

$\lambda = 0$: Interpolates data (high variance).

$\lambda \to \infty$: Simple linear regression.

Optimal λ determined by Cross Validation.



6. Local Linear Regression

Algorithm

Goal: Fit at a specific point x_0.

Selection: Find the s closest training points. 越远 → 0.

Weighting: Assign weights $K_{i0} = K(x_i, x_0)$ (decreases with distance; 0 if not in neighbor set). Fit: Minimize Weighted Least Squares: ① choice of weighting function K ② choice of local constant, linear or quadratic regression functions

$$\sum_{i=1}^{n}K_{i0}(y_i - \beta_0 - \beta_1x_i)^2$$

用 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 拟合出 x_0

Predict: $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1x_0$.

③ **Span** s 跨度

Controls flexibility of the non-linear fit. Small s: Local, wiggly fit.

Large s: Global fit (approaches linear regression). S can be 决定 by cross validation.

7. Generalized Additive Models (GAMs) 广义可加模型

Definition

General framework for multiple variables. Maintains additivity but allows non-linear functions f_i. 保持可加性

Formula: f_i can be estimated by any non linear model.

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i$$

Functions f_j can be splines, local regression, etc.

Pros & Cons model is additive in nature, can exam effect of each individual on Y, while holding all of the other variables fixed.

Pros: Automatically models non-linearities; Interpretable (examine effect of X_j holding others fixed).

Cons: Misses interactions between variables unless explicitly added (e.g., $f_{jk}(X_j, X_k)$). 排除了交互效应 / consider include

1. 分类问题概述（Classification Overview）基本定义 1

设定：输入 $X \in R^p$，输出 $Y \in \{1, \ldots, K\}$(定性变量/类别)。

分类器：函数 $G: R^p \to \{1, \ldots, K\}$。

目标：最小化误分类率 (Misclassification Error)。

$$err(G) = P(Y \neq G(X)) = E(I(Y \neq G(X)))$$

2. 分类器构建（Classifier Construction）一般框架 2

定义分类函数 $h_k(x)$(针对第 k类)。

决策规则：选择得分最高的类。 $\hat{G}(x) = \arg\max_k \hat{h}_k(x)$ by train set

$$G(x) = \text{argmax}_k h_k(x)$$

分类边界：$\{x: h_k(x) = h_l(x)\}$ 得分相等

二元分类特例（Binary Case） 3 $K = 2$，类别编码为 $\{0, 1\}$或 $\{-1,1\}$。

决策： $G(x) = I(h(x) > 0.5)$ 或 $G(x) = \text{sign}(h(x))$ 同上

边界：$\{x: h(x) = 0.5\}$或 $\{x: h(x) = 0\}$。

3. 线性回归用于分类（Linear Regression for Classification）

二分类：将响应编码为 $0/1$，拟合线性回归 $Y = X^T\beta$。预测时看 $\hat{Y} > 0.5$。

多分类问题：简单的数值编码 (1, 2, 3) 强加了类别间的顺序和距离，这是不合理的

解决方案：使用虚拟变量 (Dummy Variables) 编码 6。

创建 K个指示变量 (Indicator Response Matrix)。

对每一列拟合线性回归 $\hat{h}_k(X)$。 决策:$\hat{G}(X) = \text{argmax}_k\hat{h}_k(X)$

局限性：预测值可能 <0或 >1(Masking problem) 无法三分类 (masking problem)

4. 贝叶斯规则（Bayes Rule）

最优分类器：最小化误分类率的分类器。

公式：选择后验概率最大的类。

$$G^*(X) = \text{argmax}_kP(Y = k \mid X) = \text{argmax}_kp_k(X)$$

两种策略： 估计 $p_k(X)$(如 LDA, QDA, Logistic Regression)。后验 (tree/神网)

直接估计边界 $G^*(X)$(如 SVM) boosting/bagging 决策边界

5. 判别分析（Discriminant Analysis）

贝叶斯定理 9 $P(Y = k \mid X) = \frac{\pi_kf_k(X)}{\sum\pi_lf_l(X)}$ 假设 $f_k(X)$ 分布形状 → 概率

π_k：先验概率 $P(Y = k)$。

$f_k(X)$：第 k类的(Class-conditional density)。比较 $p_k(X)$等价于比较 $\pi_kf_k(X)$。

5.1 线性判别分析（LDA）(Discriminant Analysis)

假设：$X \mid Y = k \sim N(\mu_k, \Sigma)$(各类协方差矩阵相同)。 $f_k(X) = \frac{\exp\{-\frac{1}{2}(X-u_k)^T\Sigma^{-1}(X-u_k)\}}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}}$

判别函数 (Discriminant Function) $\log\frac{P_k(X)}{P_l(X)} > 0 \Rightarrow P_k(X) > P_l(X)$

这是一个关于 X的线性函数。

决策：分类到 $\delta_k(X)$最大的类。 $\log\frac{P_k(X)}{P_l(X)} = \log(\frac{\pi_k}{\pi_l}) - \frac{1}{2}(u_k+u_l)^T\Sigma^{-1}(u_k-u_l) + X^T\Sigma^{-1}(u_k-u_l)$

参数估计： $\hat{\pi}_k = n_k/n$。

$\hat{\mu}_k$：类均值 (Centroid)。 $\hat{\Sigma}$：合并协方差矩阵 (Pooled covariance)。

5.2 二次判别分析（QDA） 11

$\delta_k(X) = \log(\pi_k) - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + X^T\Sigma^{-1}\mu_k$

假设：$X \mid Y = k \sim N(\mu_k, \Sigma_k)$(各类协方差矩阵不同)。判别函数：

$$\delta_k(X) = \log(\pi_k) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(X - \mu_k)^T\Sigma_k^{-1}(X - \mu_k)$$

包含 X的二次项。 特点：决策边界是二次曲线。 $P(Y=k|X) = \frac{\pi_kf_k(X)}{\sum_{l=1}^{k}\pi_lf_l(X)}$

5.3 朴素贝叶斯（Naïve Bayes） $f_k(X)$ 很难估

假设：给定类别 $Y = k$，各特征 X_j相互独立。 $f_k(X) = \prod_{j=1}^{p}f_{kj}(x_j)$

优点：简化了估计任务 (只需估计单变量密度)，减少了方差 (Bias-Variance Trade-off)，在高维数据下表现通常不错。小样本. bias↑

6. 逻辑回归（Logistic Regression）6.1 二元逻辑回归 13

模型：使用逻辑函数 (Logistic/Sigmoid function) 建模 $P(Y = 1 \mid X)$。

$$p(X) = \frac{e^{\beta_0 + X^T\beta}}{1 + e^{\beta_0 + X^T\beta}}$$ $\frac{e^z}{1+e^z}$

Logit (对数几率):

$$\text{logit}(p(X)) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \boxed{\beta_0 + X^T\beta}$$ z

解释：β_j表示 X_j每增加 1 个单位，对数几率 (Log-odds) 增加 β_j(几率乘以 e^{β_j})。

参数估计 (MLE) 14 极大似然估计：最大化对数似然函数 $l(\beta)$。

$$l(\beta) = \sum[y_i\log p(x_i) + (1 - y_i)\log(1 - p(x_i))]$$

求解算法：IRLS (迭代重加权最小二乘法)，即 Newton-Raphson 方法 15。

迭代更新：

$$\beta^{new} = \beta^{old} + (X^TWX)^{-1}X^T(y - p)$$

$y^{new} = X\beta^{old} + W^{-1}(y-p)$

6.2 多元逻辑回归 16

选取一个基准类 (如第 K类)，建立 $K - 1$个 Logit 模型。

$$\log(p_k(X)/p_K(X)) = X^T\beta_k$$

6.3 Probit 模型 17

使用标准正态分布的 CDF (Φ) 代替逻辑函数。

$$p(X) = \Phi(\beta_0 + X^T\beta)$$

6.4 Logistic Regression vs LDA 18

两者都有线性的 Logit 形式。 LDA：正态 → 均值和协方差估计参数

区别：拟合过程不同。

Logistic：最大化条件似然 (MLE)。 Logit：不假设分布，用 MLE 估计

LDA：基于正态分布假设估计均值和方差 (最大化联合似然)。

7. 分离超平面（Separating Hyperplane） 19

思想：在特征空间中寻找一个超平面，将不同类别的点完全分开。

几何表示：超平面方程 $f(x) = \beta_0 + X^T\beta = 0$。 β：法向量

法向量：$\beta^* = \beta/\|\beta\|$。 normal to the surface of L

点到平面的距离：$(X^T\beta + \beta_0)/\|\beta\|$20。

决策：

$f(x) > 0 \to$Class 1 $f(x) < 0 \to$Class -1 $y^*f(x^*)$ 置信度

8. 最佳分离超平面（Optimal Separating Hyperplane 最大间隔分类器（Maximal Margin Classifier） 2

目标：找到一个超平面，不仅能分开数据，而且使离超平面最近的点（支持向量）的距离（间隔 Margin）最大化。

优化问题：

$$\max_{\beta, \beta_0} C$$

subject to

$$\frac{1}{\|\beta\|}y_i(x_i^T\beta + \beta_0) \ge C$$

等价形式 (标准形式) 22:

$$\min_{\beta, \beta_0}\frac{1}{2}\|\beta\|^2$$

subject to

$$y_i(x_i^T\beta + \beta_0) \ge 1$$

	P	N
Positive	TP	FP
Negative	FN	TN

使用 Lagrange 乘子法，转换为二次规划 (Quadratic Programming, QP) 问题求解。

9. 分类模型评估（Evaluation） 24

基础指标:误分类率 (Test Error):

$$\frac{1}{N}\sum I(\hat{y}_i \neq y_i)$$

ROC: Ture positive rate / False...

Advanced metrics: 混淆矩阵 (Confusion Matrix): TN, FP, FN, TP。

Sensitivity (Recall/TPR): TP/P(true positive rate, ability to find true positives). recall.

Specificity (TNR): TN/N(true negative rate). False positive rate = 1 − Specificity $= FP/N$

Precision (PPV): TP/P^*(proportion of true positives among predicted positives).

F1-score: $2 \times$ (Precision $\times$ Recall)/(Precision + Recall)

ROC 曲线: Plot TPR vs FPR (1-Specificity). test E: $\frac{1}{|\text{test set}|}\sum I(\hat{G}(x_i) \neq y_i)$

AUC: Area Under the ROC Curve (越大越好)。

Youden index is J = Sensitivity + Specificity − 1

1. Tree-based Methods Basics

Core Concepts Segment predictor space into a lot of simple regions.

Prediction: Use mean or mode of training observations in that region.

Properties: Region-wise constant, completely nonparametric. 非参数

2. Notations & Structure 区域常量

X: Input vector. $\in \psi$, splits of subsets of ψ into 2 descendant subsets

Nodes: Root, Terminal (Leaf $\tilde{T}$), Parent/Child (t_L, t_R). all nodes (T), node (t)

Split: s. A partition of the parent node, set of splits: S parent/child/sibling node.

3. Key Elements of Tree Construction

① Set of Questions: Binary splits. ② Goodness of Split: Criterion $\phi(s, t)$.

③ Stopping Rule: When to stop splitting. ④ Class Assignment: Assign class to leaf nodes.

3.1 Standard Set of Questions 标准问题集

分配类别 / of any node t

Depends on only one variable:

Continuous: Is $X_j <= c$? $(A = (-\infty, c])$. The splits of all p variables can be evaluated for any splits

Categorical: Is X_j in A? (A is a subset of categories). Constitute the standard set of Questions

4. Goodness of Split & Impurity

Impurity Functions $G(t)$ Let p_{tk} be proportion of class k in node t:

Misclassification Error: $1 - \max_k p_{tk}$ 不敏感 $P_{tk} = P(Y=k|t) = \frac{\sum_{i \in t}(y_i = k)}{|t|}$

Gini Index: $1 - \sum_k p_{tk}^2$ (衡量混乱程度)

Cross Entropy: $-\sum_k p_{tk}\log(p_{tk})$ 交叉熵

Criterion Maximize impurity reduction: P_L: 节点 left 的比例

$$\phi(s, t) = G(t) - p_LG(t_L) - p_RG(t_R)$$ ↑ good.

5. Stopping & Assignment Rules

Stopping Criteria 所有可能分割中，比 β 小 pre-specified threshold

① Stop if weighted gain $< \beta$: $\max_s \frac{|t|}{n}\phi(s, t) < \beta$ ⇐ 现倾向修剪枝

② Or if node is pure/small enough. may not work well, 可能有巨大收益在后

误分类误差 $P(t)$: t 中比例; $R(T) = \sum_{t \in \tilde{T}} P(t)G(t)$ 误差. R(T) is bias downward. $P(t)G(t) \ge P(t_L)G(t_L) + P(t_R)G(t_R)$

Assignment $k(t) = \arg\max_k p_{tk}$ (选可能性最大的一类)

Majority Vote (argmax $\hat{p}_{tk}$) for 0-1 loss. tree larger, misclassification error↓ overfitting!

6. Pruning & Overfitting

① 先长成 big tree ② 剪枝.

Overfitting

$R(T)$ is biased downward. Large trees = low training error, high test error (overfitting).

Minimal Cost Pruning Cost Function: 叶子节点数

λ: tuning parameter.

$$R_\lambda(T) = R(T) + \lambda|\tilde{T}|$$

迭代代价函数最小的子树.

Minimizing this balances training error $R(T)$ and complexity $|\tilde{T}|$. $\hat{T}(\lambda) = \arg\min_T R_\lambda(T) = \arg\min_T(R(T) + \lambda|\tilde{T}|)$

Optimal λ determined by Cross Validation 38.

7. Regression Tree

Impurity: Sum of Squared Errors (SSE).

$$G(t) = \sum_{i \in t}(y_i - avg(y \mid t))^2$$

Goodness of fit: $\phi(s,t) = G(t) - G(t_L) - G(t_R)$

Assignment: Node mean $\kappa(t) = avg(y \mid t)$.

8. Pros & Cons 解释

Pros: Interpretable, /very nice and intuitive graphical display

2. Automatic stepwise variable selection and complexity reduction

3. Provide an estimate of conditional class probability. 提供条件概率的估计

4. handle both categorical and ordered variables in a simple and natural way.

Cons:

High Variance: Sensitive to data changes.

Lack of smoothness.

Categorical splits: $2^{q-1} - 1$ combinations.

~~Splits based on combined variables~~

9. Bagging (Bootstrap Aggregation) 有放回抽样

Algorithm

Draw B Bootstrap samples. $(x_1^{*b}, y_1^{*b}) \ldots$

Fit tree $\hat{f}^{*b}(x)$ on each. $\hat{f}_{bag}(x) = \begin{cases}\frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) & \text{regression/class. 平均} \\ \text{mode}\{\hat{f}^*(x_1), \ldots, \hat{f}^*(x_b)\} & \text{classification 众数}\end{cases}$

Aggregate: Average (Regression) or Mode (Classification).

Intuition 42

Jensen's Inequality:

$$E(Y - \hat{f})^2 \ge E(Y - \hat{f}_{bag})^2$$

平均化可降低方差

Averaging reduces variance ("Wisdom of crowds").

Variable Importance 通过对所有决策树中每个变量引起的杂质减少量求和计算得出

Calculated by summing the impurity decrease due to each variable across all trees.

10. Random Forest bias↑ 但很小

Algorithm 和 bagging 一样

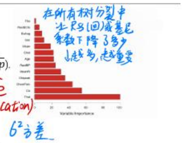
Draw B Bootstrap samples.

Grow tree, but at each split:

Randomly select p candidate predictors (usually $\sqrt{p}$).

Split using best variable in this subset.

Aggregate (average for regression / majority vote for classification)



Why it works (Variance Formula)

Bagging trees are correlated (ρ).

Variance of average of B correlated variables: σ^2 方差

$$\rho\sigma^2 + \frac{1 - \rho}{B}\sigma^2$$

Random Forest reduces ρ (Decorrelation) by random feature selection, thus reducing variance further than Bagging.

11. Boosting

Concept fitting a single large tree → overfitting / Boosting involves growing multiple trees, slow learning. but

11.1 Boosting for Regression Tree (Algorithm)

Bagging/RF 并行, boosting sequentially/adaptively use information from previously grown trees.

Init $f_0(x) = \bar{y}$, residuals $r_i = y_i$ 残差.

For $m = 1$ to M: 不抽样, 每个树都 fit on 整个数据

Fit tree h_m to residuals (x_i, r_i).

Update model: Iteratively learning weak classifiers

$$f_m(x) = f_{m-1}(x) + \alpha h_m(x)$$

Final: weighted sum of the results of each weak classifiers 加权求和.

Update residuals:

$$r_i = r_i - \alpha h_m(x_i)$$

11.2 Adaboost.M1 (Algorithm) Boosting for classification.

Init weights $w_i = 1/n$.

For $m = 1$ to M:

Fit $h_m(x)$ using weights.

Weighted error e_m.

Set

$$\alpha_m = \frac{1}{2}\log((1 - e_m)/e_m)$$

Update weights: Increase for misclassified.

$$w_{m+1,i} \propto w_{m,i}\exp(-y_i\alpha_mh_m(x_i))$$

Output sign of weighted sum.

Interpretation

Forward Stagewise Additive Model. Fits basis functions sequentially.

11.3 Regularization

Boosting test error don't rise as much as the number of iterations m ↑

Early stopping: Limit M.

Shrinkage: Use $v(0 < v < 1)$ to scale updates. 对过拟合相对免疫, 轻微

shrinkage factor $f_m(x) = f_{m-1}(x) + v \cdot \alpha_mh_m(x)$

$\sum_m \alpha_m$, M↓ v↓ ⇒ Σ↓ ⇒ train error↑, 反之亦然. v↓, 更多树 但泛化性能↑

12. Summary

Bagging, Random Forests, and Boosting are all tree-based ensemble methods, but they differ in how the trees are constructed and combined. Bagging grows many trees independently on bootstrap samples of the data, which reduces variance but often produces highly similar trees. Random forests extend bagging by additionally selecting a random subset of features at each split, decorrelating the trees and leading to better exploration of the model space. Boosting, in contrast, builds trees sequentially on the original data, with each new tree focusing on the residual errors of the previous ones and contributing gradually to improve overall prediction performance.

Kernel Trick｜ 核技巧：对于无法线性分离的数据，SVM 通过核函数将数据映射到更高维的空间，从而实现线性分离。常见核函数包括：线性核:$K(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}'$多项式核:$K(\mathbf{x},\mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^d$径向基核（RBF）:$K(\mathbf{x},\mathbf{x}') = \exp(-\gamma|\mathbf{x} - \mathbf{x}'|^2)$