Supervised Learning Methods

Alexander McLain

June 16, 2025

Case Studies

Disease Prediction

- AI-Enhanced Blood Test for Early Parkinson's Detection

Outbreak Detection

- Machine Learning-Based COVID-19 Outbreak Detection

Personalized Medicine

- MammaPrint: 70-Gene Signature for Breast Cancer Treatment Decisions

Introduction

- Today, we'll talk about supervised learning.
- Our objective is to predict outcomes for new data based on learned patterns.
- The type of supervised method to use and how to evaluate how it works depends a lot on the type of outcome.
- Continuous outcomes:
  - **Methods:** linear regression, shrinkage linear regression methods, Support vector Regression, and Neural Networks.
  - **Metrics:** mean squared prediction error, median absolute deviation
- Binary outcomes:
  - **Methods:** logistic regression, decision trees, random forests, support vector machines.
  - **Metrics:** accuracy, precision, recall, F1 score, ROC curves, and AUC.

Background: Linear Regression

- Linear Regression to model one numerical variable as a linear function of some other numeric variables.
  - the goal is to explain the variation in a variable, or to predict the value of a variable
- Two types of variables included in linear regressions
  - dependent (response, outcome) variable (Y): the variable to be predicted.
  - independent (covariates, predictors) variable ($X_1, \cdots, X_p$): the variables used to predict Y.
- The linearity is in terms of the coefficients: Y is a linear combination of independent variables plus some error.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Formal Definition of an SLR

- $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0, β_1 (unknown): intercept and coefficients, respectively.
- ϵ: random error (commonly called the residual).
- For any given value of X, $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$
- The expected SLR line: $E[Y] = \beta_0 + \beta_1 X$
- For every one unit change in X we expect Y to change by β_1 units.
- The mean value of Y given X changes by β_1 units for every 1 unit change in X.

Key Assumptions of Linear Regression (LINE)

- **Linearity:** Relationship between predictors and outcome is linear.
- **Independence:** Observations are independent from one another.
- **Normality:** Residuals (errors) are normally distributed.
- **Equal Variance:** Constant variance of errors across all levels of the predictors.

Violations can affect inference, prediction accuracy, or both.

Assessing Model Fit: Residual Analysis

- Residuals = observed − predicted values

$$e_i = Y_i - \hat{Y}_i$$

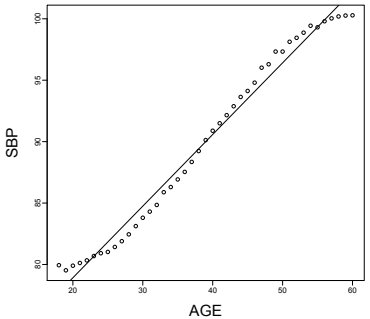
- Plots to examine:
  - Residuals vs. fitted: check linearity & homoscedasticity
  - Q-Q plot: check normality
  - Residuals vs. time/order: check independence
- Look for:
  - No clear pattern in plots
  - Residuals centered around 0

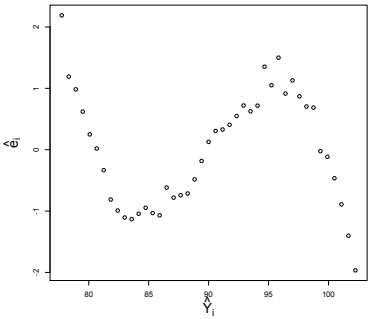
Line

- Assumption: there is a linear relationship between Y and the X's. I.e.,
  - The expected value of the residual is zero for all combinations of X's
  - $E(e_i) = 0$.
- How to check
  - Plot Y_i vs. X_i.
  - Plot e_i vs. $\hat{Y}_i$

Line

Example Violation:





Corrective actions:

- Fit a different regression model.
- Add quadratic of cubic components.
- Transformation of Y and/or X

Normality

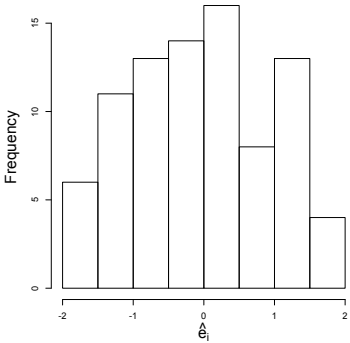
- Assumption: The residuals are normally distributed.
- How to check
  - Histogram of the e_i's.
  - Normal probability plot of e_i.
  - Tests of normality on e_i.
  - Outlier tests.

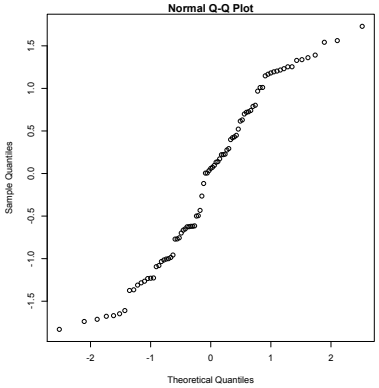
Normality

Example Violation:

Corrective actions:

- Examine outliers to determine if they are contaminated
- Transformation of Y and/or X

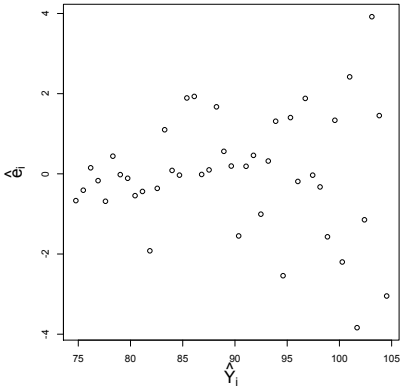




Equal Variance

- **Assumption:** the variance of the residuals is equal for all X's.
- How to check
  - Plot Y_i vs. X_i.
  - Plot e_i vs. $\hat{Y}_i$

  Example Violation $\Longrightarrow$



- **Corrective action:**
  - Transformation of Y.
  - Patterned covariance model

Visual Checks for Assumptions

- **Linearity:** Scatterplots of Y vs. X, or residuals vs. fitted values
- **Independence:** Time series plots (look for patterns)
- **Normality:** Histogram or Q-Q plot of residuals
- **Equal Variance:** Residuals vs. fitted plot (look for "funnel" shape)

Residual analysis is your main tool to diagnose model issues.

The Problem of Model Selection

- In linear regression, we often have many potential predictors.
- **Model selection** = deciding which subset of variables to include.
- Why it matters:
  - Too few predictors ⇒ underfitting
  - Too many predictors ⇒ overfitting
- Goal: balance complexity and prediction accuracy

Data Example

POLAR (Predicting Outcomes of Language Rehabilitation in Aphasia) trial

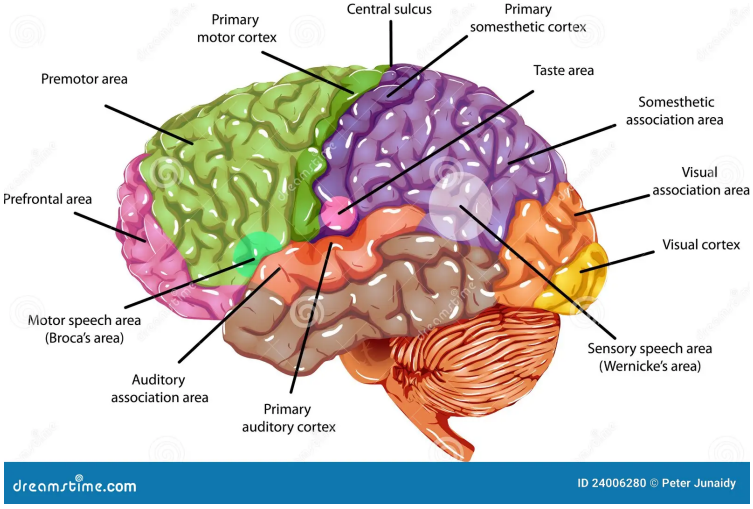
- A total of 107 stroke patients with chronic aphasia (speech disorder) were randomized to one of two treatment arms.
- Neuroimaging data on where their stroke occurred is available on $\geq 5 \times 10^6$ voxels
- Main outcome is the Western Aphasia Battery (WAB).
- Goals of the study:
  1. see which treatment arm was the most effective,
  2. predict a person's WAB score based on their neuroimaging data,
  3. determine which areas of the brain with damage are related to WAB, and
  4. determine which areas of the brain with damage are related to treatment efficacy.

Multiple Linear Regression

- One dependent variable Y.
- p independent variables:
  - X_j = proportion of voxels damaged in Region Of Interest (ROI) j.
- The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  or $E(\boldsymbol{Y}) = \boldsymbol{X}'\boldsymbol{\beta}$.
- β_0: intercept;
- $\beta_j, j = 1, \cdots, p$: regression coefficients.
- $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$


Primary motor cortex
Central sulcus
Primary somesthetic cortex
Premotor area
Taste area
Somesthetic association area
Visual association area
Prefrontal area
Visual cortex
Motor speech area (Broca's area)
Sensory speech area (Wernicke's area)
Auditory association area
Primary auditory cortex
dreamstime.com
ID 24006280 © Peter Junaidy

What $\boldsymbol{X}$ to use?

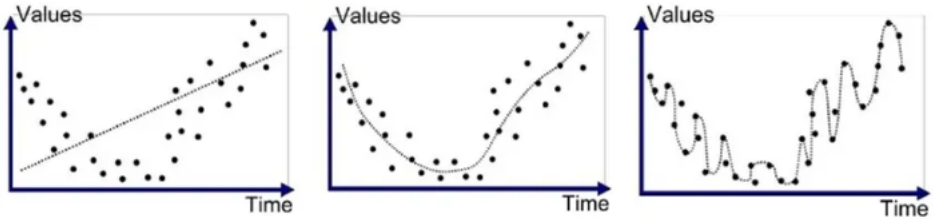
- Recall, that X_j = proportion of voxels damaged in ROI j.
- How do we define the regions? Some options:
  1. **Harvard-Oxford Atlas,** Number of ROIs (i.e., p): 48
  2. **Automated Anatomical Labeling (AAL) Atlas,** Number of ROIs: 116
  3. **Brainnetome Atlas,** Number of ROIs: 246
  4. **Schaefer Atlas,** Number of ROIs: Varies (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000)
  5. **Voxel level,** $p > 10^5$.

Model Selection Goals:

To make reasonable predictions or estimations, we need

- accuracy $\rightarrow$ on average, what we estimate is equal to what we expect.
  - The predicted AQ score is equal to the average AQ score for all groups of people
- precision $\rightarrow$ small variation in prediction/estimation
  - The predicted AQ score is close to the true AQ score for all groups of people.

Underfitting vs. Overfitting



Underfitting vs. Overfitting

- Underfitting occurs when important regressors are left out of the model. Costs:
  - deficient models (i.e., missing patterns).
  - misinterpretations of variable relationships.
- Overfitting occurs when all important regressors are in the model, but some unimportant ones are, too. Costs:
  - unneeded complexity and increased variance of the predicted values.
  - widened confidence and prediction intervals.

Common Model Selection Approaches

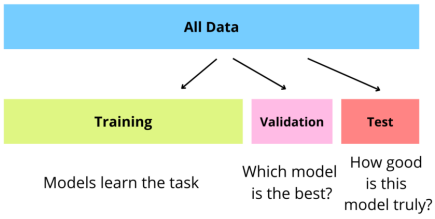
- **Best Subset Selection:** Try all combinations
- **Stepwise Selection:** Add or remove variables step-by-step
- **Penalized Methods:** Add penalty to control model size
  - LASSO (L1 penalty), Ridge (L2 penalty)

Evaluation Criteria

- **Training Error:** How well the model fits existing data
- **Test Error:** How well it predicts new data
- **Cross-Validation:** Estimate performance on unseen data
- **Information Criteria:**
  - AIC: balances fit and complexity
  - BIC: stronger penalty for complexity

Introduction to Validation

- Validation involves splitting the data into training, validation, and test sets.
- **Training Set:** The portion of data used to train the model.
- **Validation Set:** The portion of data used to tune model parameters and prevent overfitting. It's used for intermediate evaluation during model training.
- **Test Set:** The portion of data used to assess the final performance of the model after training and validation.



Introduction to Cross-Validation

- **Definition**: Cross-validation (CV) is a statistical method used to estimate the performance of machine learning models.
- **Purpose**: It helps in assessing how a model generalizes to an independent dataset.
- Why Use Cross-Validation?
  - **Prevents Overfitting:** Ensures model's robustness.
  - **Provides Reliable Estimates:** Offers a very good estimate of model performance.
  - **Optimizes Hyperparameters**: Helps choose parameters that can't be estimated (tuning parameters, e.g., the number of variables in a model).
  - **Accuracy Measures:** Any type can be used.

Basic Concepts

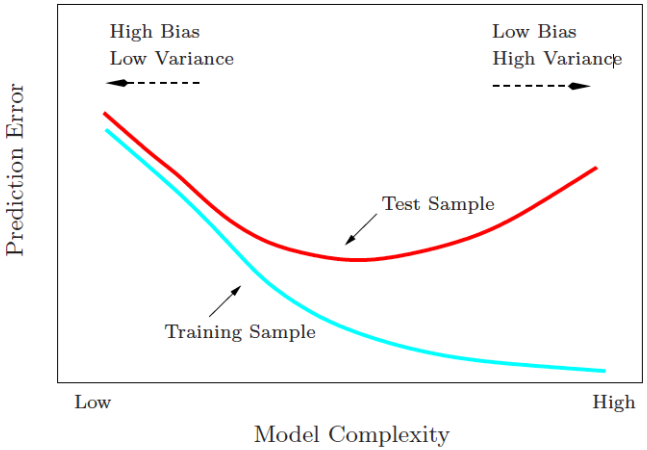
- **Divide Data:** Split data into training and validation (or test) sets.
- **Multiple Iterations:** Perform the split multiple times.
- **Aggregate Results:** Average the performance metrics.
- Types of CV:
  - K-Fold Cross-Validation
  - Leave-One-Out Cross-Validation (LOO)
  - Stratified K-Fold Cross-Validation
  - Time Series Split (for time-dependent data)

Cross-Validation Details

- Steps of K-fold cross-validation:
  1. Split data into K subsets (folds).
  2. Train/Estimate on K-1 folds and va
     on the remaining fold.
  3. Repeat K times.
  4. Average the results.

- With LOO, each person is their own fold.

Bias-Variance Tradeoff



What $\boldsymbol{X}$ to use?

- Recall, that X_j = proportion of voxels damaged in ROI j.
- How do we define the regions? Some options:
  1. **Harvard-Oxford Atlas,** Number of ROIs (i.e., p): 48
  2. **Automated Anatomical Labeling (AAL) Atlas,** Number of ROIs: 116
  3. **Brainnetome Atlas,** Number of ROIs: 246
  4. **Schaefer Atlas,** Number of ROIs: Varies (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000)
  5. **Voxel level,** $p > 10^5$.

Example: Predicting AQ with brain images (with 5-fold CV)

1. Put the people into 5 groups.
   Let $G_i = k$ if person i is in group k
2. Then for $k = 1, 2, 3, 4, 5$:
   - 2.1 Use $k - 1$ folds to fit a linear model to every brain atlas.
   - 2.2 Using the fitted linear models, predict AQ for every person in the kth group, i.e., the test fold.

   - 2.3 Get the total error for the test group for each brain atlas. For brain atlas 'a' this would be:

$$SSE_k^a = \sum_{i;G_i=k} (Y_i - \hat{Y}_i^a)^2$$

where $\hat{Y}_i^a$ is the predicted AQ when brain atlas 'a' is used.

Example: Predicting AQ with brain images (cont.)

3. Get the total error over all folds:

$$SSE^a = \sum_{k=1}^{5} SSE_k^a, \quad SSE^b = \sum_{k=1}^{5} SSE_k^b, \quad SSE^c = \sum_{k=1}^{5} SSE_k^c, \quad \text{etc.}$$

4. Use the ROI that has the lowest SSE.

Classification

Classification Introduction

- We'll now switch from continuous to binary (yes/no) types of outcomes.
- For example, the Wisconsin Breast Cancer Dataset (WBCD) was created by Dr. William H. Wolberg at the University of Wisconsin Hospitals, Madison.
- These data are used to predict whether a breast cancer tumor is benign or malignant based on various cell features.

Classification Introduction

- We'll now switch from continuous to binary (yes/no) types of outcomes.
- For example, the Wisconsin Breast Cancer Dataset (WBCD) was created by Dr. William H. Wolberg at the University of Wisconsin Hospitals, Madison.
- These data are used to predict whether a breast cancer tumor is benign or malignant based on various cell features.
- **Examples of cell features:**
  - Radius (mean of distances from the center to points on the perimeter),
  - Texture (standard deviation of gray-scale values),
  - Perimeter, Area, etc.

  Each of these features is calculated for three different metrics: mean, standard error, and the "worst" or largest value.

Classification Introduction

- Classification is the process of predicting the class label of a given input based on training data that contains input-output pairs.
- It involves building a model that learns from the training data and can predict the class of new, unseen data.
- Linear regression cannot be used for classification.
- Some Common Algorithms: logistic regression, decision trees, K-Nearest Neighbors (KNN), Naive Bayes, Neural Networks, Gradient Boosting Machines.

What is Logistic Regression?

- A type of regression used when the outcome is **binary** (e.g., Yes/No, 0/1)
- Predicts the **probability** that the outcome is 1 (success)
- Output is between 0 and 1, but not linear in X
- Example: Will a patient develop a disease (Yes/No)?

The Logistic Model

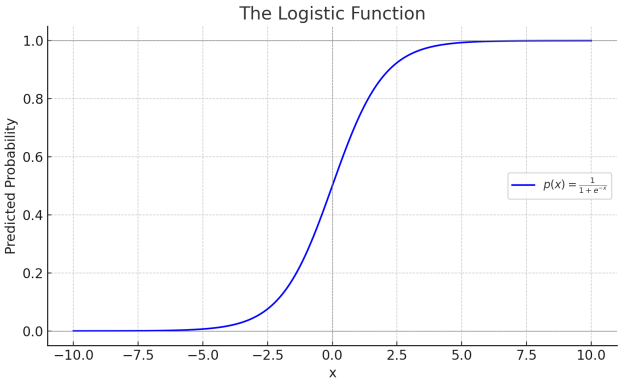
- The model uses the **logit** transformation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- This means:
  - p = probability of success
  - Left-hand side = log-odds of success
- The inverse of this function is the **logistic curve**

The Logistic Curve

- S-shaped curve
- As predictor increases, the probability approaches 0 or 1
- Ensures outputs stay in the [0,1] range



Interpreting Coefficients

- Each β represents the change in the **log-odds** for a 1-unit increase in the predictor
- $\exp(\beta) =$ **odds ratio**
- Example:
  - If $\beta_1 = 0.7$, then $\exp(0.7) \approx 2.01$
  - $\rightarrow$ A 1-unit increase in x_1 doubles the odds of success

When to Use Logistic Regression

- Outcome is binary (0/1)
- Predictors can be continuous, binary, or categorical
- Examples in public health:
  - Predicting disease presence (yes/no)
  - Smoking status, vaccine response, hospital readmission
- Assumes independent observations and a linear relationship in the log-odds

Logistic Regression: Predicting Smoking Status from Demographic and Health Factors

Goal: Predicting Smoking Status from Demographic and Health Factors.

- Handout

Model Selection in Logistic Regression

- The same questions about model selection come up.
- For example, which of the cell features should be used for our $\boldsymbol{x}$? Which summary measure should we use?
- AIC and BIC are two common measures of the "quality" of a model, which are general and can be used in almost any method.

Introduction

- The above question on model fit are usually used for **predictive models.**
- However, in public health we are often more interested in determining the association between two variables.
- For example, what is the association between the use of multivitamins and mortality?

Introduction

- Rarely is it reasonable to talk about the association of two variables without consideration of the impact of other variables.
- You will see a variety of labels for different etiologic and statistical phenomena that might be occurring.
- For example
  - Confounding
  - Direct Effect.
  - Indirect Effect.
  - Mediation
  - Modification/moderation
  - Interaction

Confounding

- Confounding reflects the causal association between variables in the population under study
- A confounder is an extraneous variable that satisfies the following criteria:
  - It is a risk factor for the study disease.
  - It is associated with the study exposure, but is not a consequence of that exposure.
  - The association with disease must occur in the absence of exposure.
- A confounder is a risk factor for the study disease whose "control" in some appropriate way will reduce (or remove) bias in estimating the exposure–disease relationship.

Confounding

- If we are interested in estimating the OR, then we could obtain a biased estimate of the OR if a confounder is not adjusted for (e.g., examining MV–MI without adjusting for age or smoking).
  - OR could be positively or negatively biased.
- How do we control/identify confounding?
  - Must have prior knowledge on potential confouders
  - Is the adjusted estimate of the OR $\approx$ to the crude estimate of the OR?
  - People often use the 10% rule (i.e., a 10% change in the OR is a sign of confounding).

Berkley Gender Bias Case:

- Data from: "Sex Bias in Graduate Admissions: Data from Berkeley," *Science* 187: 398-403; 1975.
- In 1973 2,681 men and 1,835 women applied to graduate school, with 44% of men and 35% of women being admitted.
- A crude analysis finds Crude OR $= (1276/1835)/(1486/2681) = 1.25$ with 95% CI $(1.20, 1.32)$.
- This difference is statistically significant (i.e., not due to chance).

Berkley Gender Bias Case:

The admission rates and RR by department

	Men		Women		
Depart.	Applicants	Admitted	Applicants	Admitted	OR_i
A	825	62%	108	82%	0.90
B	560	63%	25	68%	0.99
C	325	37%	593	34%	1.08
D	407	35%	375	34%	1.02
E	191	23%	393	26%	0.88
F	373	7%	341	6%	1.09

- The Summary OR is 0.97 and insignificant.
- The apparent association (OR=1.25) was due to confounding.

Mediation

- Can be thought of as a special case of a potential confounder being part of a causal pathway.
- if $E \Rightarrow M \Rightarrow D$ then we say M is a mediator.

For example:

- TAAG (Trial for Activity among Adolescent Girls)
- Interventions (E) effected physical activities (D).
- The E may have changed self-image or self-efficacy (M).
- The could be a relationship between M and D.
- Important to understand why physical activity increased.

Effect modification

- When we observe that the *OR* are non-constant across strata, we say there is **statistical interaction**.
- Interaction (a statistical term) is related to effect modification (an epi term).
- If the effect of E on D varies with C we say that there C is an effect modifier.
  - **Effect modification:** a variable that differentially (positively and negatively) modifies the observed effect of a risk factor on disease status. Present if different groups (of C) have different risk estimates.
  - Effect modification is related to the biology of disease, not just a data observation (what makes it different from interaction).
- An obvious example of interaction is breast or prostate cancer.

Effect modification

Why study effect modification? Why do we care?

- to define high-risk subgroups for preventive actions,
- to increase precision of effect estimation by taking into account groups that may be affected differently,
- to increase the ability to compare across studies that have different proportions of effect-modifying groups, and
- to aid in developing a causal hypotheses for the disease.

Effect Modification vs Confounding

Interesting tidbit:

- If a variable C is a confounder, then the stratum specific estimates of the $\widehat{OR}_i$ will mostly be on one side of the crude estimate of the $\widehat{OR}$.
- If a variable C is an effect modifier, then the crude estimate of the $\widehat{OR}$ will be a "weighted average" of the stratum specific estimates of the $\widehat{OR}_i$.
- **Note:** we are comparing the crude OR **NOT** the adjusted estimate of the OR.

What to do under effect modification/confounding?

1. If a variable is a confounder, you just have to adjust for it:

$$\beta_0 + \beta_1 E + +\beta_2 C$$

2. If a variable is an effect modifier, it should go into your regression equation as an **interaction term**

$$\beta_0 + \beta_1 E + +\beta_2 C + \beta_3 EC$$

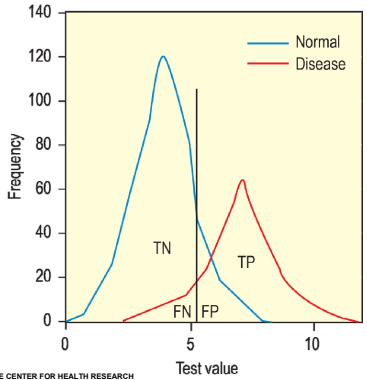
Measuring of predictive ability

- We will now take a slight detour to discuss other ways of quantifying the predictive ability of a model.
- Some of these methods will take more background than others.
- In general, they are all trying to answer the following:

  Ten new subjects walk into the room (all were not in the data used to estimate $\hat{\beta}$), which model is going to give me the best estimate of the predictive probability (i.e., $\hat{\pi}_i$) for those ten subjects?

Sensitivity and specificity

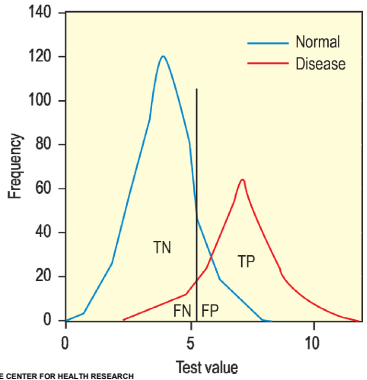
- Imagine a study evaluating a new test that screens people for a disease $\{0, 1\}$.
- The test outcome can be positive (predicting that the person has the disease) or negative (predicting that the person does not have the disease).
- The test results for each subject may or may not match the subject's actual status.



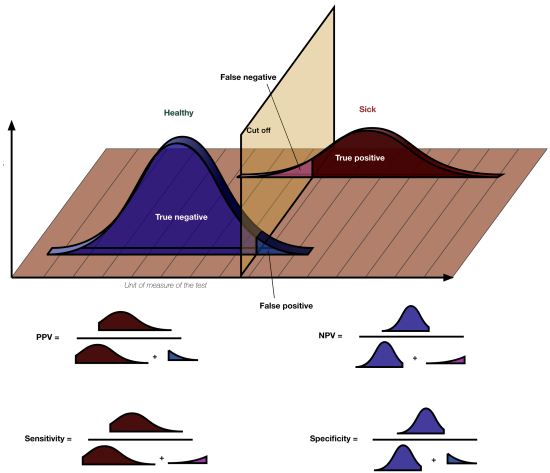
Sensitivity and specificity

In this setting, there are 4 things that could happen:

- **True positive (TP):** Sick people correctly diagnosed as sick
- **False positive (FP):** Healthy people incorrectly identified as sick
- **True negative (TN):** Healthy people correctly identified as healthy
- **False negative (FN):** Sick people incorrectly identified as healthy



Sensitivity, specificity, PPV, NPV

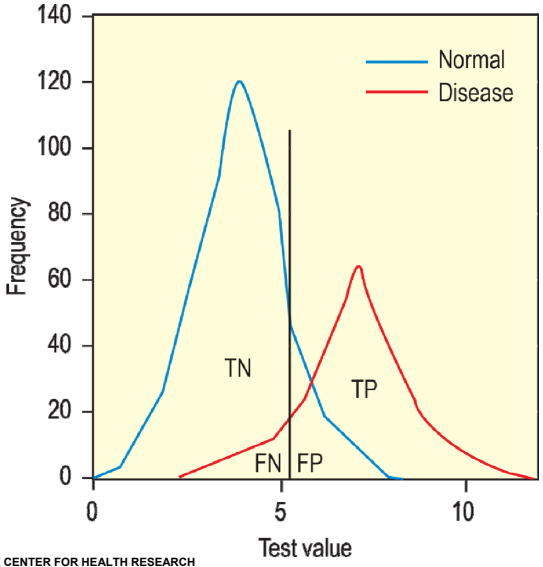


Estimating sensitivity and specificity from a logistic model

- To estimate sensitivity and specificity from a logistic model we we'll use the values of $\hat{\pi}_i$ as the results of the "test."
- Let's assume that $Y = 1$ if the person has the disease, so that a high value of $\hat{\pi}_i$ is indicative that the person will have the disease.
- Then we can pick a value, say π_0, such that the result of the test is

$$T_i = \begin{cases} +, & \text{if } \hat{\pi}_i > \pi_0 \\ -, & \text{if } \hat{\pi}_i \leq \pi_0 \end{cases}$$

- So all people with a predicted value of at least π_0 will have a test equal to "positive".



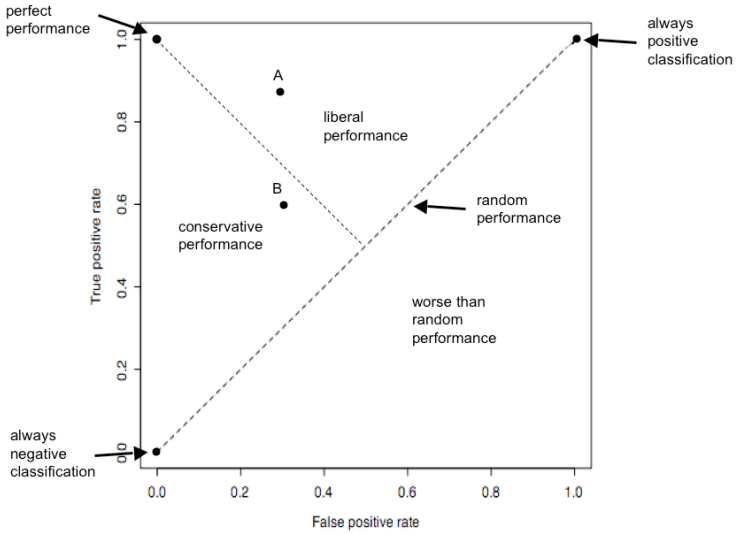
Receiver Operating Characteristic (ROC) Curves

- Receiver Operating Characteristic (ROC) Curves are a method for measuring the predictive ability of a model.
- ROC curves look at the $Sens(\pi_0)$ and $1 - Spec(\pi_0)$ for all values of π_0.

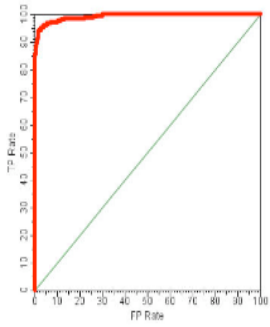
- When $\pi_0 = 0$ all tests are positive, $Sens(0) = 1$ and $1 - Spec(1) = 1$.
- When $\pi_0 = 1$ all tests are negative, $Sens(1) = 0$ and $1 - Spec(1) = 0$.

Motivation for ROC Curves

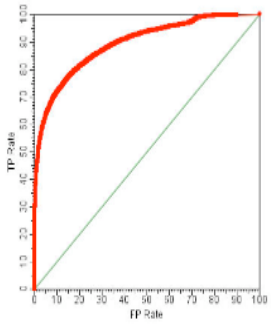
- For example, assume we have 100 diseased and 100 not diseased in our sample.
- Suppose that we list all 200 people from smallest to largest $\hat{\pi}_i$.
- Suppose that when I increase π_0 from 0.4 to 0.5 an additional 20 people have T^-.
- If the test is worthless I would expect 10 of those people to be diseased and 10 not diseased.
- In general, (If the test is worthless) as π_0 increases, 50% of the people are diseased and 50% are not diseased.


perfect performance
always positive classification
A
liberal performance
B
random performance
conservative performance
worse than random performance
always negative classification
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

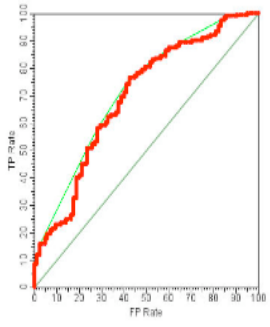
ROC curve examples



Almost perfect prediction



Good prediction



Poor prediction

Quantifying ROC curve performance

- Notice that the models that predict better have larger area under the ROC curve (AUC).
- AUC is the main way to quantify an ROC curve.
- The AUC is equal to the probability that the model will rank a randomly chosen diseased individual higher than a randomly chosen healthy individual.

$$\Pr(\hat{\pi}_i > \hat{\pi}_j | Y_i = 1 \text{ and } Y_j = 0)$$

AUC values

- We will interpret the AUC in terms of the discriminative ability of the model.
- That is, how well does the model discriminate between diseased and non-diseased individuals?

$\widehat{AUC}$	Interpretation of discriminative ability
< 0.5	worse than expected by chance
$= 0.5$	no discrimination
$0.7 - 0.8$	acceptable discrimination
$0.8 - 0.9$	excellent discrimination
> 0.9	outstanding discrimination

Logistic Regression: COVID-19 Diagnosis

Goal: Distinguish COVID-19 from other respiratory illnesses

- Logistic regression model based on 400 patients
- 15 features, achieved 98.8% sensitivity, 97.3% specificity
- **Source:** PMC9277749

CART Introduction

- Suppose we have two variables X_1 and X_2, and we are trying to classify group status.
- Tree based methods consider breaking the X_1 and X_2 space into blocks.

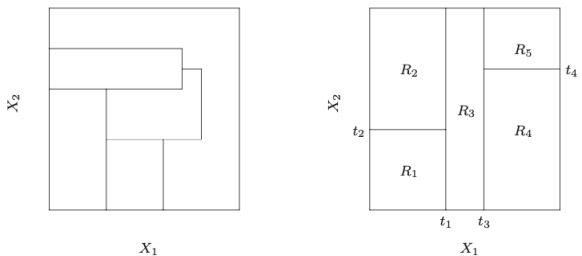


Figure: CART Example from ESL II (page 306).

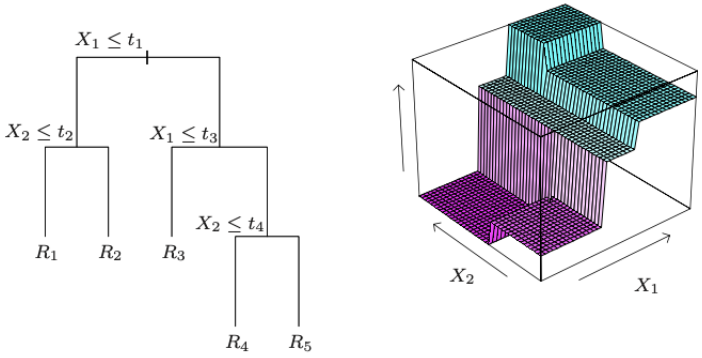


Figure: CART Example from ESL II (page 306).

- The model predict population C_m for $(X_1, X_2) \in R_m$.

Growing Trees

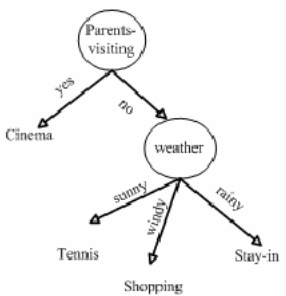
- We now turn to the question of how to grow a regression tree.
- Our data consists of p inputs and a response, for each of N observations: that is, (x_i, y_i) for $i = 1, 2, \ldots, N$, with $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
- The algorithm decides what variable to split on and the split point.
- We split the data to maximize the differences in the outcome for the two splits.
- For example, split such that the sample means in the two groups are as different as possible.

How far to grow?

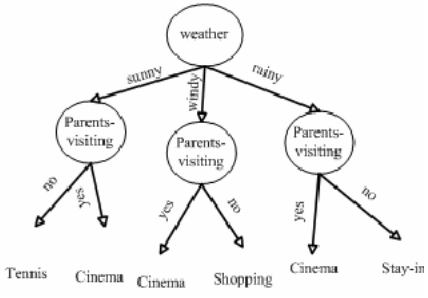
- Tree size is a tuning parameter governing the model's complexity.
- The preferred strategy is to grow a large tree T_0, stopping the splitting process only when some minimum node size (say 5) is reached.
- The large tree is pruned using **cost-complexity pruning**.

Pruned Tree

a). Prunned decision tree produced by RBDT-1.



b). Prunned decision tree produced by AQDT-1 & AQDT-2.



CART: Predicting Influenza Risk

Goal: Identify flu patients from symptoms + demographics

- CART used to generate fast, interpretable triage tools
- Effective alternative to lab testing in resource-limited settings
- **Source:** PMC5034457

CART: COVID-19 Mortality Risk

Goal: Predict severe COVID-19 outcomes in hospital patients

- Dataset: 5,000+ hospitalized patients
- CART classified mortality risk $\rightarrow$ clinical triage
- **Source:** Int J Emerg Med 2024

Summary Discussion

- Logistic regression: good for interpretation and risk scoring
- CART: good for decision support and threshold-based tools
- Consider both accuracy and interpretability
- **Discussion:** Which would you use for your own project?