Mathematical Statistics: Foundations

Content

- Mathematical Statistics in AI
- Probability Review
- Statistic Methods
- Linear Algebra
- Optimizations Introduction (gradient descent, hyperparameters)
- Calculus

Mathematical Statistics in AI

- **PROBABILITY**
  - Basic Rules and Axioms
  - Random Variables
  - Bayes' Theorem
  - Distributions: Binomial, Bernoulli, Poisson, Exponential, Gaussian
  - Conjugate Priors
- **LINEAR ALGEBRA**
  - Vectors
  - Matrices
  - Eigenvalues & Eigenvectors
  - Principal Component Analysis
  - Singular Value Decomposition
- **CALCULUS**
  - Functions
  - Scalar Derivative
  - Gradient
  - Vector and Matrix Calculus
  - Gradient Algorithms

Probability Review

- **MOTIVATION**
  - The agent needs reason in an uncertain world

  - Uncertainty can be due to
    - Noisy sensors (e.g., temperature, GPS, camera, etc.)
    - Imperfect data (e.g., low resolution image)
    - Missing data (e.g., lab tests)
    - Imperfect knowledge (e.g., medical diagnosis)
    - Exceptions (e.g., all birds fly except ostriches, penguins, birds with injured wings, dead birds, ...)
    - Changing data (e.g., flu seasons, traffic conditions duringrush hour, etc.)

      ...

  - The agent still must act (e.g., step on the breaks, diagnose a patient, order a lab test, ...)

Probability Review

- **TENTATIVE PLAN**
  - Probability background
  - Classification
    - Naïve Bayes, logistic regression, neural networks
    - Maximum likelihood estimation, Bayesian estimation, gradient optimization, backpropagation
  - Decision-making
    - Episodic decision-making, Markov decision processes, multi-armed bandits
    - Value of information, Bellman equations, value iteration, policy iteration, UCB1, -greedy
  - Reinforcement learning
    - Prediction, control, Monte-Carlo methods, temporal difference learning, Sarsa, Q-learning

Probability Review

- **SOME EXERCISES**
  - In a class, 70% of the hardworking students got an A. John got an A. What is the probability that John is a hardworking student?

  - You design a Covid test with the following behavior
    - $P(+ \mid \text{covid}) = 0.95$; $P(- \mid \text{covid}) = 0.05$
    - $P(+ \mid {\sim}\text{covid}) = 0.10$; $P(- \mid {\sim}\text{covid}) = 0.90$
    - John takes the test, and the result is +. What is the probability that John has covid?

  - In a town, 70% of the hospitalized are vaccinated. Do the vaccines provide any protection against hospitalization?

  - P(toothache | cavity) = 0.75. P(cavity | toothache) = ?

Probability Review

- **RANDOM VARIABLES**
  - Pick variables of interest
    - Medical diagnosis
      - Age, gender, weight, temperature, LT1, LT2, ...
    - Loan application
      - Income, savings, payment history, ...
    - Earlier examples
      - Grad student, Grade, Covid, Test result, Ache, X-Ray
  - Every variable has a domain
    - Binary (e.g., True/False)
    - Categorical (e.g., Red/Green/Blue)
    - Real-valued (e.g., 97.8)
  - Possible world
    - An assignment to all variables of interest

Probability Review

- **PROBABILITY MODEL**
  - A probability model associates a numerical probability P(w) with each possible world *w*
    - P(*w*) sums to 1 over all possible worlds

  - An event is the set of possible worlds where a given predicate is true
    - Roll two dice
      - ❑ The possible worlds are (1,1), (1,2), ..., (6,6); 36 possible worlds
      - ❑ Predicate = two dice sum to 10
      - ❑ Event = {(4,6), (5,5), (6,4)}
    - Toothache and cavity
      - ❑ Four possible worlds: (t, c), (t, ~c), (~t, c), (~t, ~c)
      - ❑ Some worlds are more likely than others
      - ❑ Predicate can be anything about these variables: $t \wedge c$,t,$t \vee \sim c$,

Probability Review

- **AXIOMS OF PROBABILITY**
    - The probability P(a) of a proposition a is a real number between 0 and 1
    - P(true) = 1, P(false) = 0
    - $P(a \lor b) = P(a) + P(b) – P(a \land b)$

Probability Review

- **$P(\neg a)$**
    - $P(a \lor \neg a) = P(a) + P(\neg a) - P(a \land \neg a)$
    - $P(\text{true}) = P(a) + P(\neg a) - P(\text{false})$
    - $1 = P(a) + P(\neg a) - 0$
    - $P(\neg a) = 1 - P(a)$
    - Intuitive explanation:
        - The probability of all possible worlds is 1
        - Either a or a holds in one world
        - The worlds that a holds and the worlds that a holds are mutually exclusive and exhaustive

Probability Review

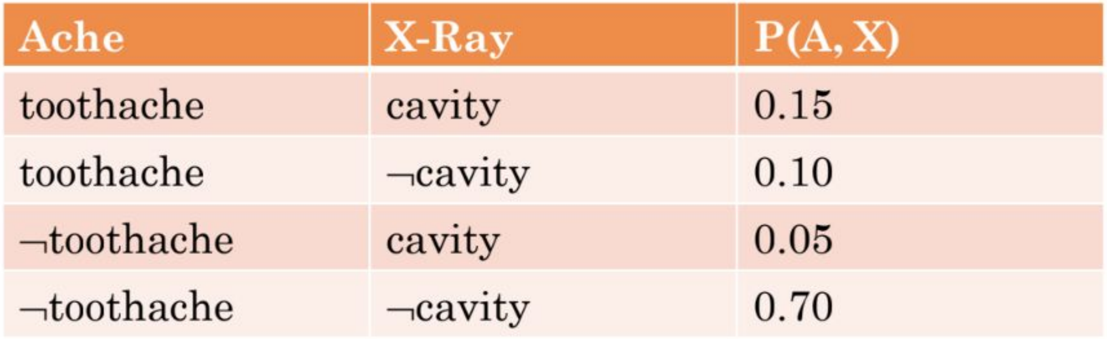
- **RANDOM VARIABLES – NOTATION**
  - Capital: X: a variable
  - Lowercase: x: a particular value of X
  - $Val(X)$: the set of values X can take
  - Bold Capital: X: a set of variables
  - Bold lowercase: x: an assignment to all variables in X
  - $P(X=x)$ will be shortened as $P(x)$
  - $P(X=x \wedge Y=y)$ will be shortened as $P(x, y)$

Probability Review

- **JOINT DISTRIBUTION**
  - We have n random variables, V1, V2, ..., Vn
  - We are interested in the probability of a possible world, where
    - V1 = v1, V2 = v2, ..., Vn = vn
  - P(V1, V2, ..., Vn) associates a probability for each possible world the **joint distribution.**
  - How many entries are there, if we assume the variables are all binary?

Probability Review

- **TOOTHACHE EXAMPLE**



Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

Probability Review

- **PRIOR AND POSTERIOR**
  - Prior probability
    - Probability of a proposition in the absence of any other information
    - E.g., P(V1, V3, V5)
  - Conditional/posterior probability
    - Probability of a proposition given another piece of information
    - E.g., P(V2, V3 | V5 = T, V7 = F)
    - P(A | B) = P(A B) / P(B)

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

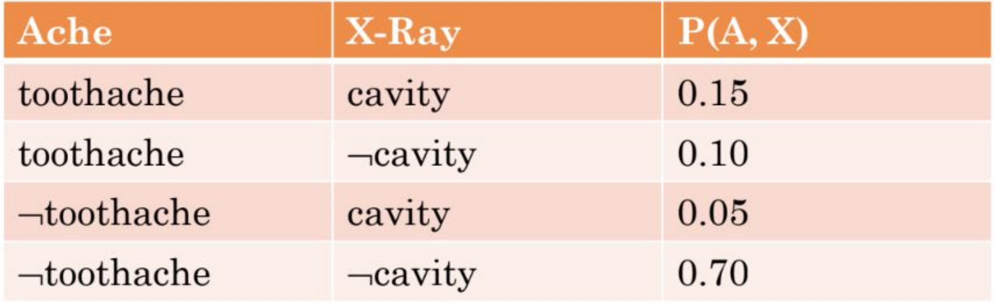
Probability Review

- **MARGINALIZATION**
  - Given a distribution over n variables, you can calculate the distribution over any subset of the variables by summing out the irrelevant ones

  - For example
    - Probability of a proposition given another piece of information
    - Given P(A, B, C, D)
      - ❏ Calculate P(A)
      - ❏ P(A, C)
      - ❏ ... (any subset)

Probability Review

- **LET’S ANSWER A FEW QUERIES**



Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

  - P(cavity) = ?
  - P(cavity) = ?
  - P(toothache) = ?
  - P(toothache) = ?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

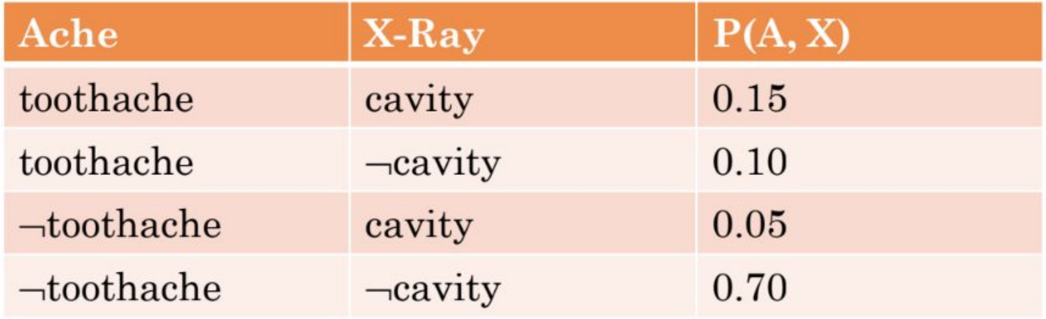
Probability Review

- **CONDITIONAL DISTRIBUTION**
  - $$P(A,B,C \mid D,E,F,G) = \frac{P(A,B,C,D,E,F,G)}{P(D,E,F,G)}$$

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **LET'S ANSWER A FEW QUERIES**
  - P(cavity | toothache) = ?
  - P(cavity | toothache) = ?
  - P(cavity | toothache) = ?
  - P(cavity | toothache) = ?
  - P(toothache | cavity) = ?
  - P(toothache | cavity) = ?
  - P(toothache | cavity) = ?
  - P(toothache | cavity) = ?



Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

Probability Review

- **BAYES' RULE**
  - $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$

Probability Review

- **BAYES' RULE**
  - $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$
  - Example use
    - P(cause|effect) = P(effect|cause)*P(cause) / P(effect)
  - Why is this useful?
    - Because in practice it is easier to get probabilities for P(effect|cause) and P(cause) than for P(cause|effect)
      - ❑ E.g., P(disease|symptoms) = P(symptoms|disease)*P(disease) / P(symptoms)
      - ❑ It is easier to know what symptoms diseases cause. It is harder to diagnose a disease given symptoms.

Probability Review

- **BAYES' RULE**
    - Can we compute P($\alpha|\beta$) from P($\beta|\alpha$)?

Probability Review

- **CLASS EXAMPLE**
  - In a class, 70% of the hardworking students got an A. John got an A. What is the probability that John is a hardworking student?

  - Possible worlds: 4
    - <h, a>, <h, ~a>, <~h, a>, <~h, ~a>

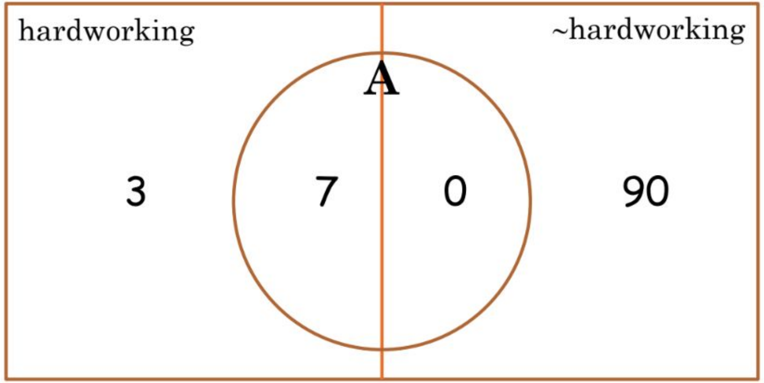
  - Let’s say there are 100 students in a class
  - Let’s say 10 of them work hard (h), 90 do not (~h)

  - Probability of a randomly picked student being hardworking
    - P(h) = 0.1
  - We are told that 70% of the hardworking students got an A.
    - P(a|h) = 0.7
    - 7 hardworking students got an A; 3 did not get an A.

  - What is P(h|a) = ?

Probability Review

- **VERY DIFFICULT CLASS**

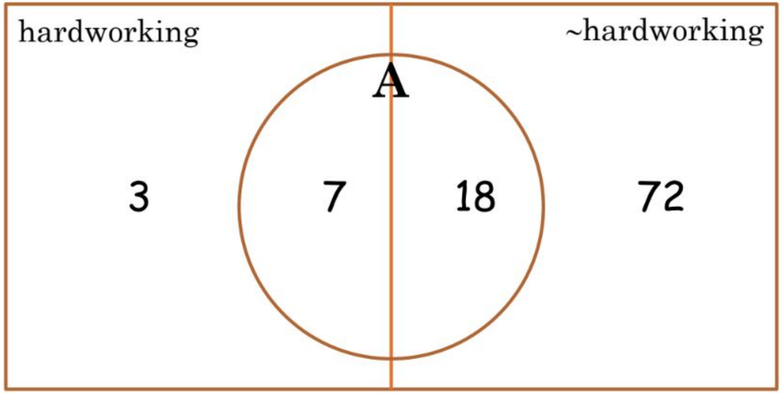


P(h | a) = ?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **MEDIUM DIFFICULT CLASS**

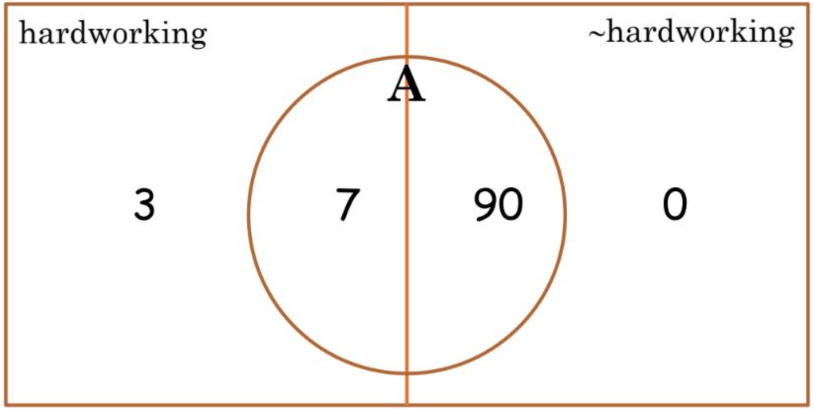


P(h | a) = ?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **WEIRD CLASS**



P(h | a) = ?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **CHAIN RULE**

  $P(X1, X2, X3, ..., Xk) =$

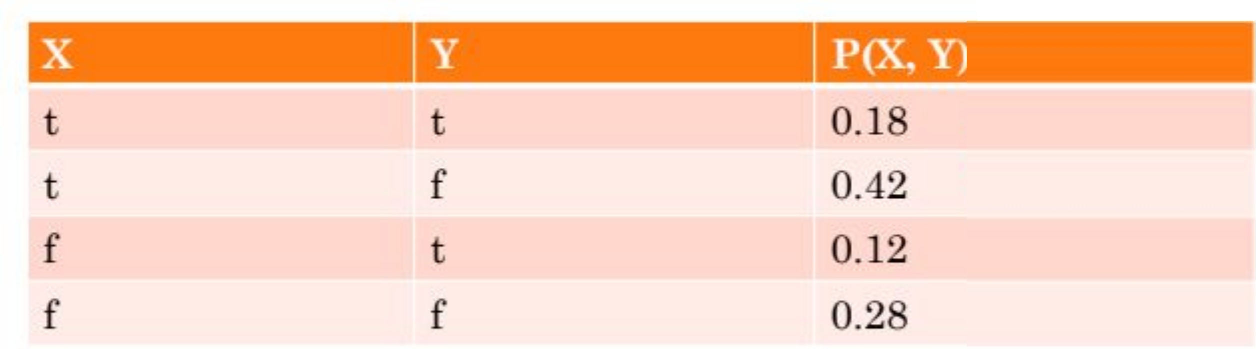
  - $P(X1)\ P(X2|X1)\ P(X3|X1,X2)...\ P(Xk\ |X1, X2, X3, ..., Xk\text{-}1)$
    - or
  - $P(X2)\ P(X1|X2)\ P(X3|X1,X2)...\ P(Xk\ |X1, X2, X3, ..., Xk\text{-}1)$
    - or
  - $P(X2)\ P(X3|X2)\ P(X1|X3,X2)...\ P(Xk\ |X1, X2, X3, ..., Xk\text{-}1)$
    - or
  - Pick an order, then P(first)P(second|first)P(third|first,second)...P(last|all_previous)

Probability Review

- **MARGINAL INDEPENDENCE**
  - An event α is independent of event β in P, denoted as P $\models \alpha \perp \beta$, if
    - $P(\alpha|\beta) = P(\alpha)$, or
    - $P(\beta) = 0$
  - Proposition: A distribution P satisfies $\alpha \perp \beta$ if and only if
    - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
    - Can you prove it?
  - Corollary: $\alpha \perp \beta$ implies $\beta \perp \alpha$

Probability Review

- **MARGINAL INDEPENDENCE**



X	Y	P(X, Y)
t	t	0.18
t	f	0.42
f	t	0.12
f	f	0.28

Is $X \perp Y$?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **CONDITIONAL INDEPENDENCE**
  - Two events are independent given another event

  - An event α is independent of event β given event in P, denoted as P $\models (\alpha \perp \beta \mid \gamma)$, if
    - $P(\alpha|\beta, \gamma) = P(\alpha| \gamma)$, or
    - $P(\beta, \gamma) = 0$

  - Proposition: A distribution P satisfies $\alpha \perp \beta \mid$ if and only if
    - $P(\alpha, \beta|\gamma) = P(\alpha|\gamma)\, P(\beta|\gamma)$

Probability Review

- **NUMBER OF PARAMETERS**
  - Assuming everything is binary
  - P(V1) requires
    - 1 independent parameter
  - P(V1, V2, ..., Vn) require
    - 2n-1 independent parameters
  - P(V1|V2) requires
    - 2 independent parameters
  - P(V1,V2, ..., Vn | Vn+1, Vn+2, ..., Vn+m) requires
    - 2m (2n-1) independent parameters

Probability Review

- **CONTINUOUS SPACES**
  - Assume X is continuous and Val(X) = [0,1]
  - If you would like to assign the same probability to all real numbers in [0, 1], what is, for e.g., P(X=0.5) = ?

Probability Review

- **PROBABILITY DENSITY FUNCTION**
  - We define probability density function, p(x), a non-negative integrable function, such that ( ) 1Val X p x dx =

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx$$

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **UNIFORM DISTRIBUTION**
  - A variable X has a uniform distribution over [a,b] if it has the PDF

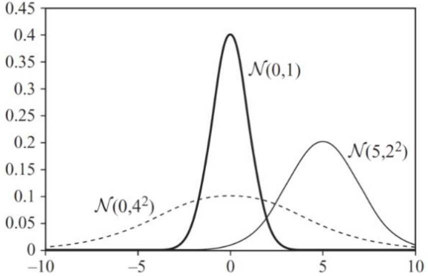
$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

Check and make sure that p(x) integrates to 1.

Probability Review

- **GAUSSIAN DISTRIBUTION**
  - A variable X has a Gaussian distribution with mean and variance 2, if it has the PDF

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Can p(x) be ever greater than 1?

Probability Review

- **CONDITIONAL PROBABILITY**
  - We want P(Y|X=x) where X is continuous, Y is discrete
  - P(Y|X=x) = P(Y,X=x) / P(X=x)
    - What's wrong with this expression?
  - Instead, we use the following expression

$$P(Y \mid X = x) = \lim_{\varepsilon \to 0} P(Y \mid x - \varepsilon \le X \le x + \varepsilon)$$

Probability Review

- **CONDITIONAL PROBABILITY**
  - We want P(Y|X=x) where X is continuous, Y is discrete
  - How would you represent it?

Probability Review

- **EXPECTATION**

$$E_P\left[X\right]=\sum_x xP(x)$$

$$E_P\left[X\right]=\int_x xp(x)dx$$

$$E_P\left[aX+b\right]=aE_P\left[X\right]+b$$

$$E_P\left[X+Y\right]=E_P\left[X\right]+E_P\left[Y\right]$$

$$E_P\left[X\mid y\right]=\sum_x xP(x\mid y)$$

What about E[X*Y]?

Mustafa Bilgic (2023). CS 581 – Advanced Artificial Intelligence – Illinois Institute of Technology

Probability Review

- **VARIANCE**

$$Var_P[X] = E_P\left[\left(X - E_P[X]\right)^2\right]$$

$$Var_P[X] = E_P\left[X^2\right] - \left(E_P[X]\right)^2$$

Can you derive the second expression using the first expression?

$$Var_P[aX + b] = a^2 Var_P[X]$$

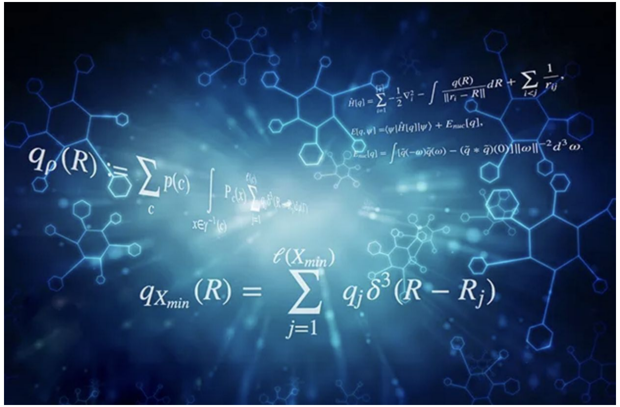
What is Var[X+Y]?

Probability Review

- **UNIFORM AND GAUSSIAN DISTRIBUTION**
  - If XN(,2), then E[X] = , Var[X] = 2
  - What about the expectation and variance of a uniform distribution?

Statistic Methods

- **Importance of Statistics in Data Science and AI**
  - Statistics is the grammar of science, especially in fields like Computer Science, Physical Science, and Biological Science
  - Statistical knowledge helps leverage data insights and understand algorithms beyond implementation.
- **Prerequisites:**
  - Basic mathematical skills (algebra, basic calculus)
  - Logical thinking for problem-solving
  - Computer literacy (basic knowledge of using computers and the internet)


$q_\rho(R) := \sum_c p(c)$
$q_{X_{min}}(R) = \sum_{j=1}^{\ell(X_{min})} q_j \delta^3(R - R_j)$

Statistic Methods

- **Key Statistical Concepts**
  - Random variables
  - Mean, variance, and standard deviation
  - Covariance and correlation
  - Probability distribution functions (PDFs)
  - Bayes’ Theorem
  - Linear Regression and Ordinary Least Squares (OLS)
  - Gauss-Markov Theorem
  - Confidence intervals

Statistic Methods

- **Key Statistical Concepts**
    - Hypothesis testing
    - Statistical significance
    - Type I & Type II Error
    - Statistical tests (Student’s t-test, F-test, 2-Sample T-Test, 2-Sample Z-Test, Chi-Square Test)
    - p-value and its limitations
    - Inferential Statistics
    - Central Limit Theorem & Law of Large Numbers
    - Dimensionality reduction techniques (PCA, FA)

Statistic Methods

- **COVARIANCE AND CORRELATION**
  - **Covariance**
    - Covariance measures how much the movement in one variable predicts the movement in a corresponding variable.
    - Covariance quantifies the co-variability of two variables around their respective means.
    - It reveals whether two variables move in the same or opposite directions.
    - Like variance, which focuses on the variability of a single variable around its mean, covariance assesses the relationship between two variables.
  - **Correlation**
    - Correlation refers to any statistical relationship between two random variables or bivariate data. Specifically, it measures the degree to which a pair of variables are linearly related.
    - A correlation coefficient is a number between -1 and 1 that quantifies the strength and direction of the relationship between variables.

Statistic Methods

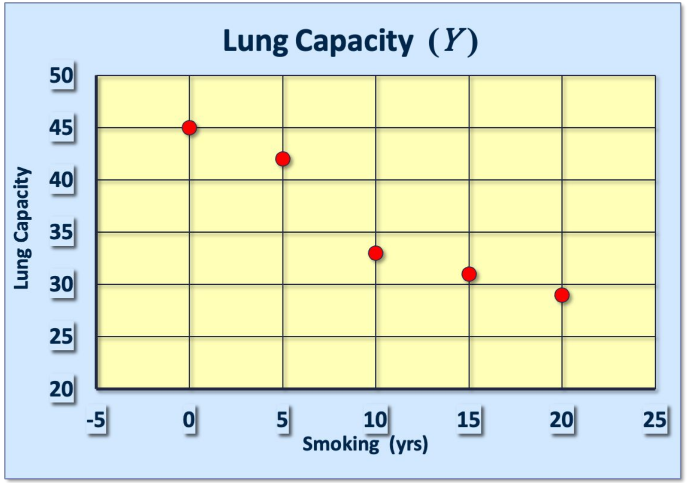
- **COVARIANCE**
  - **Example**
    - Investigate relationship between cigarette smoking and lung capacity.
    - Data: sample group response data on smoking habits, and measured lung capacities, respectively.

N	Cigarettes (*X*)	Lung Capacity (*Y*)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

Smoking and Lung Capacity Data

Statistic Methods

- **COVARIANCE**
  - Investigate relationship between cigarette smoking and lung capacity.
    - Observe that as smoking exposure goes up, corresponding lung capacity goes down
    - Variables covary inversely
    - Covariance and Correlation quantify relationship
  - Variables that covary inversely, like smoking and lung capacity, tend to appear on opposite sides of the group means.
  - Average product of deviation measures extent to which variables covary, the degree of linkage between them.



Smoking and Lung Capacity Data

Statistic Methods

- **COVARIANCE**
  - The Sample Covariance
    - Similar to variance, for theoretical reasons, average is typically computed using (N -1), not N . Thus,

[illegible]

Statistic Methods

- **COVARIANCE**
  - Calculating Covariance

Cigs (*X*)	**Lung Cap (*Y*)**
0	45
5	42
10	33
15	31
20	29

$\bar{X} =$ 10	$\bar{Y} =$ 36

Statistic Methods

- **COVARIANCE**
  - ○ Calculating Covariance

Evaluation yields, [illegible]

Cigs (X)	$(X-\bar{X})$	$(X-\bar{X})(Y-\bar{Y})$	$(Y-\bar{Y})$	*Cap (Y)*
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		∑= -215		

Statistic Methods

- **COVARIANCE**
    - Covariance under Affine Transformation

[illegible] and [illegible]

[illegible]

Evaluating, in turn, gives,

$$S_{LM} = \frac{1}{N-1}\sum_{i=1}^{N}(\Delta l)_i(\Delta m)_i$$

Evaluating further,

$$S_{LM} = \frac{1}{N-1}\sum_{i=1}^{N}(\Delta l)_i(\Delta m)_i$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}a(\Delta x)_i\, c(\Delta y)_i$$

$$= ac\frac{1}{N-1}\sum_{i=1}^{N}(\Delta x)_i(\Delta y)_i$$

$$\boxed{\therefore\ S_{LM} = acS_{xy}}$$

Statistic Methods

- **COVARIANCE**
  - Covariance under Affine Transformation

[illegible]

[illegible]

Evaluating, in turn, gives,

$$S_{LM} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\Delta l\right)_i\left(\Delta m\right)_i$$

Evaluating further,

$$S_{LM} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\Delta l\right)_i\left(\Delta m\right)_i$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}a\left(\Delta x\right)_i c\left(\Delta y\right)_i$$

$$= ac\frac{1}{N-1}\sum_{i=1}^{N}\left(\Delta x\right)_i\left(\Delta y\right)_i$$

$$\boxed{\therefore\ S_{LM} = acS_{xy}}$$

Statistic Methods

- **CORRELATION COEFFICIENT (PEARSON)**
  - Like covariance, but uses Z-values instead of deviations. Hence, invariant under linear transformation of the raw data.

$$r \sum^{N} = z_x z_y \frac{1}{}$$

  - Alternative (common) Expression

$$r \frac{S_{xy}}{}$$

Statistic Methods

- **CORRELATION COEFFICIENT (PEARSON)**
  - Computational Formula 1

$$r_{XY} = \sum_{i=1}^{N} X_i \left(\sum_{i=1}^{N} Y_i \right.$$

  - Computational Formula 2

$$r_{xy} = \quad Y - \sum X \sum Y \qquad N \sum XY$$

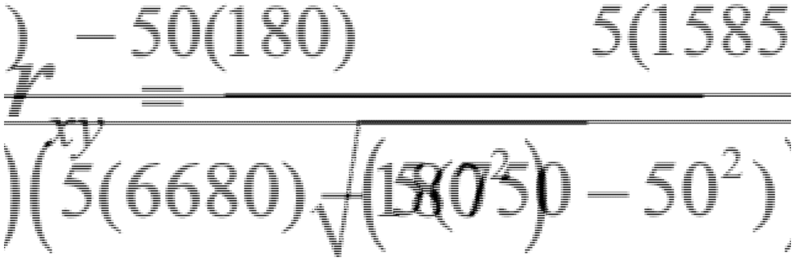
Statistic Methods

- **CORRELATION COEFFICIENT (PEARSON)**
  - Example: table for Calculating r_{xy}

	Cigs (X)	X^2	XY	Y^2	Cap (Y)
	0	0	0	2025	45
	5	25	210	1764	42
	10	100	330	1089	33
	15	225	465	961	31
	20	400	580	841	29
$\sum=$	50	750	1585	6680	180

Statistic Methods

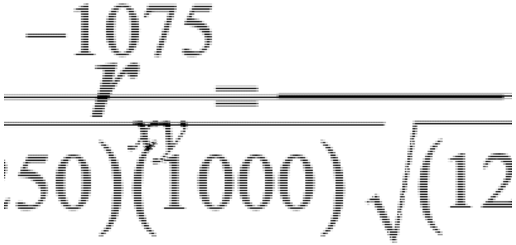
- **CORRELATION COEFFICIENT (PEARSON)**
  - Conclusion
    - r_{xy} = -0.96 implies almost certainty smoker will have diminish lung capacity.
    - Greater smoking exposure implies greater likelihood of lung damage



$$) - 50(180) \qquad 5(1585$$

$$r_{xy} =$$

$$)(5(6680) \sqrt{(5(750) - 50^2)}$$



$$-1075$$

$$r_{xy} =$$

$$50)(1000) \sqrt{(12}$$

Statistic Methods

- **BAYES' THEOREM**
  - Basic Probability Formulas
    - Product rule

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

    - Sum rule

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

    - Bayes theorem

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

    - Theorem of total probability, if event Ai is mutually exclusive and probability sum to 1

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Statistic Methods

- **BAYES' THEOREM**
  - Given a hypothesis h and data D which bears on the hypothesis:

$$P(D \mid h)P(h)$$

$$P(h \mid D) =$$

- **P(h):** independent probability of h: prior probability
- **P(D):** independent probability of D
- **P(D|h):** conditional probability of D given h: likelihood
- **P(h|D):** conditional probability of h given D: posterior probability

Statistic Methods

- **BAYES' THEOREM**
  - **Example: Does patient have cancer or not?**

    A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 99% of the cases and a correct negative result in only 95% of the cases. Furthermore, only 0.03 of the entire population has this disease.

    - What is the probability that this patient has cancer?
    - What is the probability that he does not have cancer?
    - What is the diagnosis?

Statistic Methods

- **BAYES' THEOREM**
  - **Maximum A Posterior**
    - Based on Bayes Theorem, we can compute the Maximum A Posterior (MAP) hypothesis for the data
    - We are interested in the best hypothesis for some space H given observed training data D.

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h \mid D) \\
&= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D \mid h) P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}}\, P(D \mid h) P(h)
\end{aligned}
$$

- H: set of all hypothesis.
- Note that we can drop P(D) as the probability of the data is constant (and independent of the hypothesis).

Statistic Methods

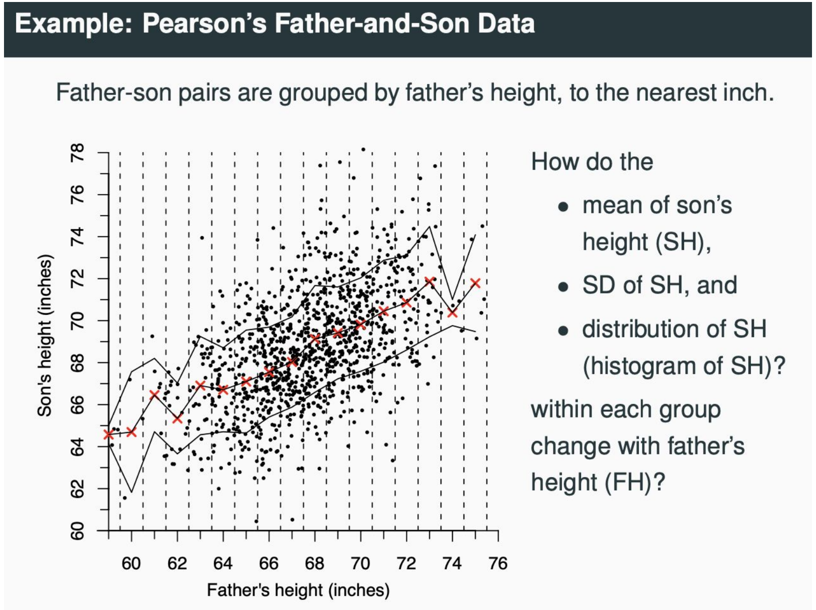
- **BAYES’ THEOREM**
  - **Maximum Likelihood**
    - Now assume that all hypotheses are equally probable a priori, i.e., P(hi) = P(hj) for all hi, hj belong to H.
    - This is called assuming a uniform prior. It simplifies computing the posterior:

$$h_{ML} = \arg\max P(D|h)$$

    - This hypothesis is called the maximum likelihood hypothesis.

Statistic Methods

- **LINEAR REGRESSION**
  - Simple Linear regression models



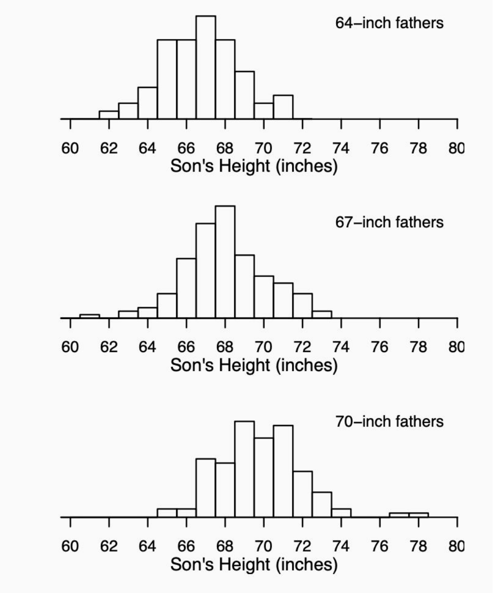
Example: Pearson's Father-and-Son Data

Father-son pairs are grouped by father's height, to the nearest inch.

How do the

- mean of son's height (SH),
- SD of SH, and
- distribution of SH (histogram of SH)?

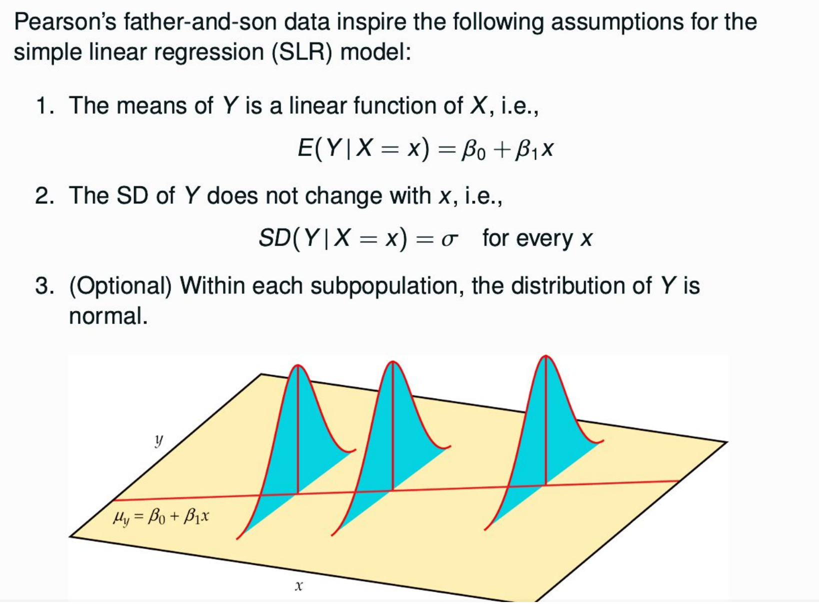
within each group change with father's height (FH)?



Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **LINEAR REGRESSION**
  - Simple Linear regression models



Pearson's father-and-son data inspire the following assumptions for the simple linear regression (SLR) model:

1. The means of Y is a linear function of X, i.e.,

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

2. The SD of Y does not change with x, i.e.,

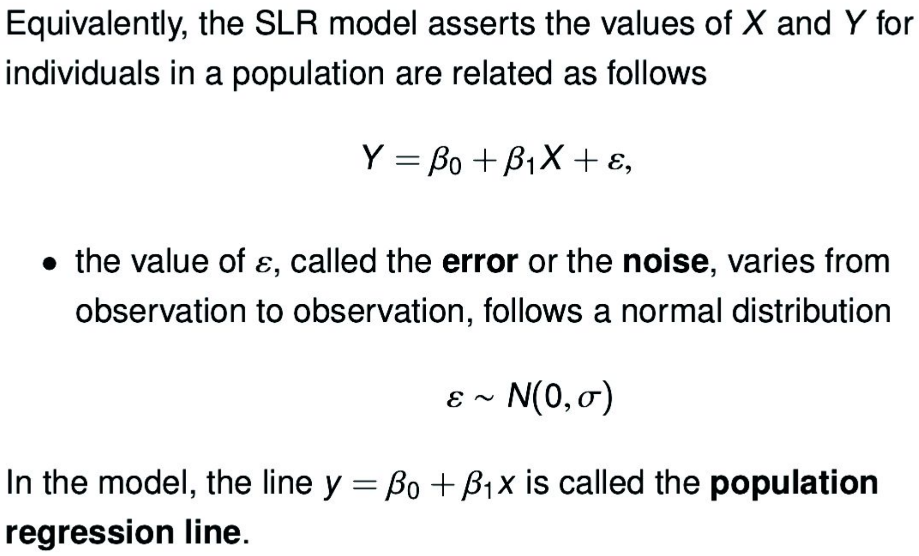
$$SD(Y \mid X = x) = \sigma \quad \text{for every } x$$

3. (Optional) Within each subpopulation, the distribution of Y is normal.

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **LINEAR REGRESSION**
  - Simple Linear regression models



Equivalently, the SLR model asserts the values of X and Y for individuals in a population are related as follows

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

- the value of ε, called the **error** or the **noise**, varies from observation to observation, follows a normal distribution

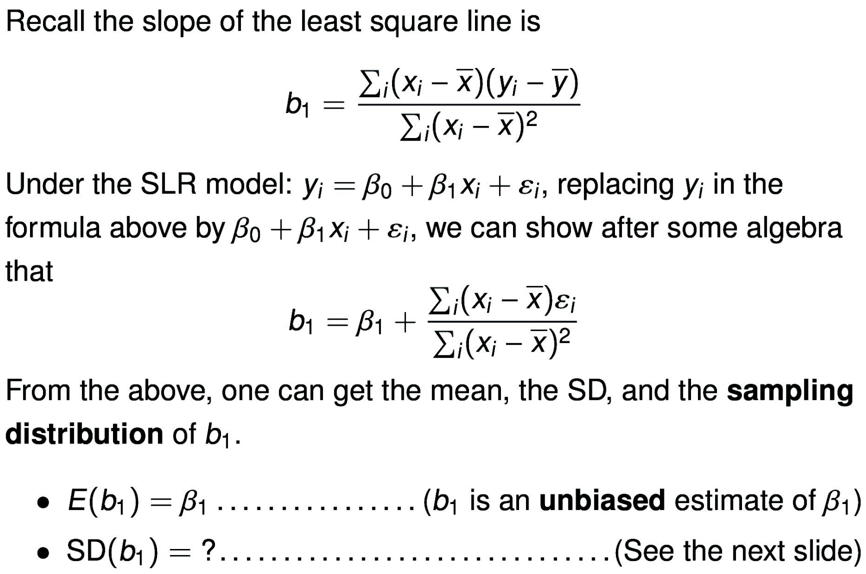
$$\varepsilon \sim N(0, \sigma)$$

In the model, the line $y = \beta_0 + \beta_1 x$ is called the **population regression line**.

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **Inference for Simple Linear Regression Models**
  - How Close Is b1 to β1?



Recall the slope of the least square line is

$$b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, replacing y_i in the formula above by $\beta_0 + \beta_1 x_i + \varepsilon_i$, we can show after some algebra that

$$b_1 = \beta_1 + \frac{\sum_i (x_i - \overline{x})\varepsilon_i}{\sum_i (x_i - \overline{x})^2}$$

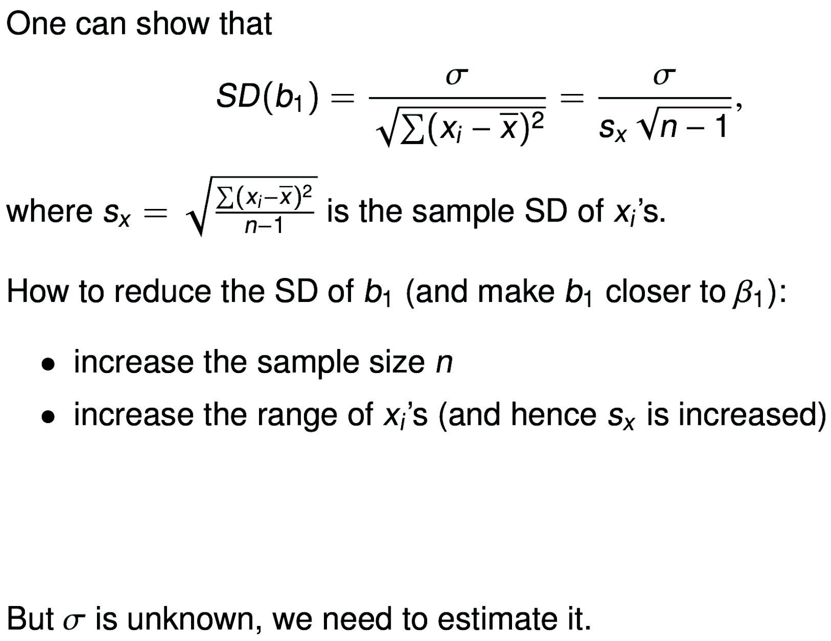
From the above, one can get the mean, the SD, and the **sampling distribution** of b_1.

- $E(b_1) = \beta_1$ .... (b_1 is an **unbiased** estimate of β_1)
- $\text{SD}(b_1) = ?$ .... (See the next slide)

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **Inference for Simple Linear Regression Models**
  - Variability of b1



One can show that

$$SD(b_1) = \frac{\sigma}{\sqrt{\sum(x_i - \overline{x})^2}} = \frac{\sigma}{s_x\sqrt{n-1}},$$

where $s_x = \sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}$ is the sample SD of x_i's.

How to reduce the SD of b_1 (and make b_1 closer to β_1):

- increase the sample size n
- increase the range of x_i's (and hence s_x is increased)

But σ is unknown, we need to estimate it.

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **Inference for Simple Linear Regression Models**
  - Estimate of σ

We want to estimate σ, SD of the error ε_i.

- An intuitive estimate of σ is the sample SD of the *errors* ε_i

$$\widehat{\sigma} = \sqrt{\frac{\sum(\varepsilon_i - \overline{\varepsilon})^2}{n-1}} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

However, this is not possible β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with b_0 and b_1 and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \widehat{y}_i$$

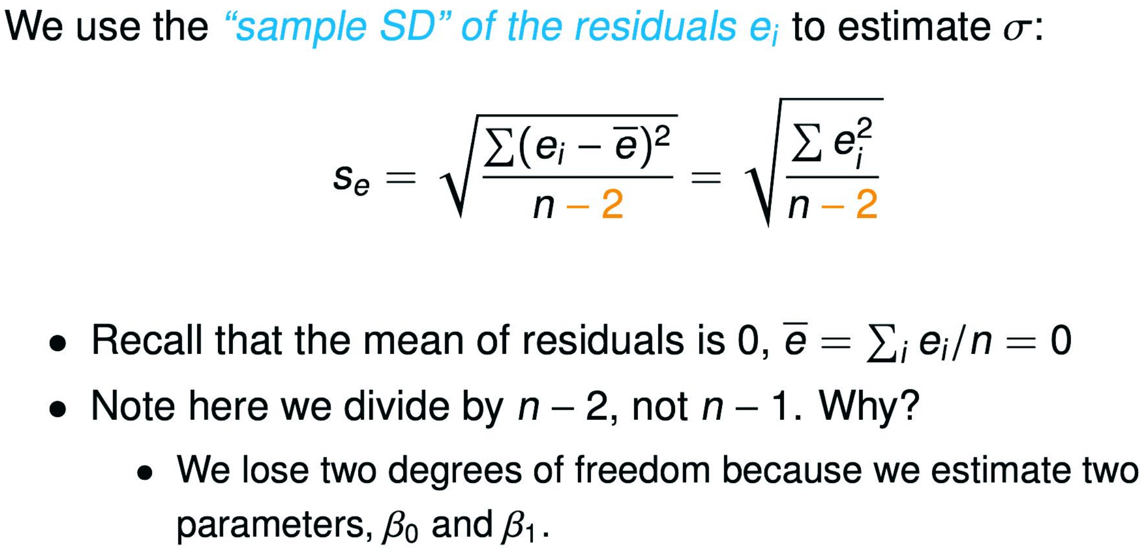
We use the “sample SD” of the residuals e_i to estimate σ:

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

Statistic Methods

- **Inference for Simple Linear Regression Models**
  - Estimate of σ



We use the *“sample SD” of the residuals e_i* to estimate σ:

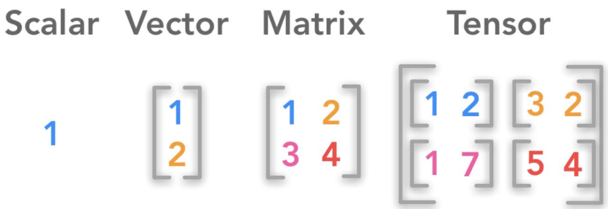
$$s_e = \sqrt{\frac{\sum(e_i - \overline{e})^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

- Recall that the mean of residuals is 0, $\overline{e} = \sum_i e_i/n = 0$
- Note here we divide by $n - 2$, not $n - 1$. Why?
  - We lose two degrees of freedom because we estimate two parameters, β_0 and β_1.

Huang, Y. (Year). STAT 220 Lecture Slides Inference for Linear Regression. Department of Statistics, University of Chicago.

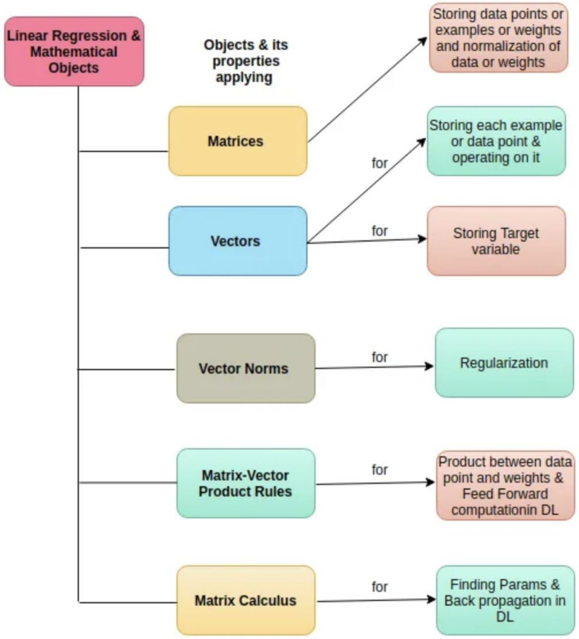
Linear Algebra

- **INTRODUCTION**
  - Linear algebra is a sub-field of mathematics concerned with vectors, matrices, and linear transforms.
  - It is a key foundation to the field of machine learning, from notations used to describe the operation of algorithms to the implementation of algorithms in code.
  - Mathematical objects in linear algebra
    - Scalar
    - Vector
    - Matrix
    - Tensors



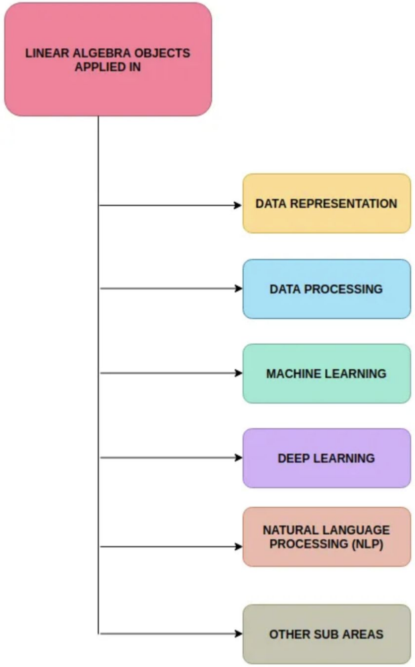
Linear Algebra

- **How Linear Algebra is applying in AI**



Linear Regression & Mathematical Objects
Objects & its properties applying
Matrices
Storing data points or examples or weights and normalization of data or weights
Vectors
for
Storing each example or data point & operating on it
for
Storing Target variable
Vector Norms
for
Regularization
Matrix-Vector Product Rules
for
Product between data point and weights & Feed Forward computationin DL
Matrix Calculus
for
Finding Params & Back propagation in DL

Linear Algebra Objects, properties and usages in ML and DL



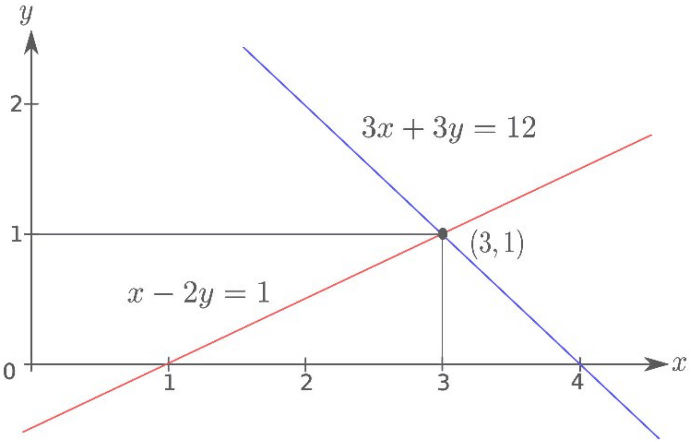
Linear Algebra Objects applying in these areas of AI

Shafi. (2022, January 30). *Linear Algebra- How uses in Artificial Intelligence ?* Analytics Vidhya.

Linear Algebra

- **EXAMPLE**
  - A classic problem is to solve systems of linear equations like
  - How about the dimension of the vector space increases beyond two?

$$3x + 3y = 12$$
$$x - 2y = 1$$



Linear Algebra

- **EXAMPLE**
  - ○ A classic problem is to solve systems of linear equations like
  - ○ How about the dimension of the vector space increases beyond two? –> Matrices

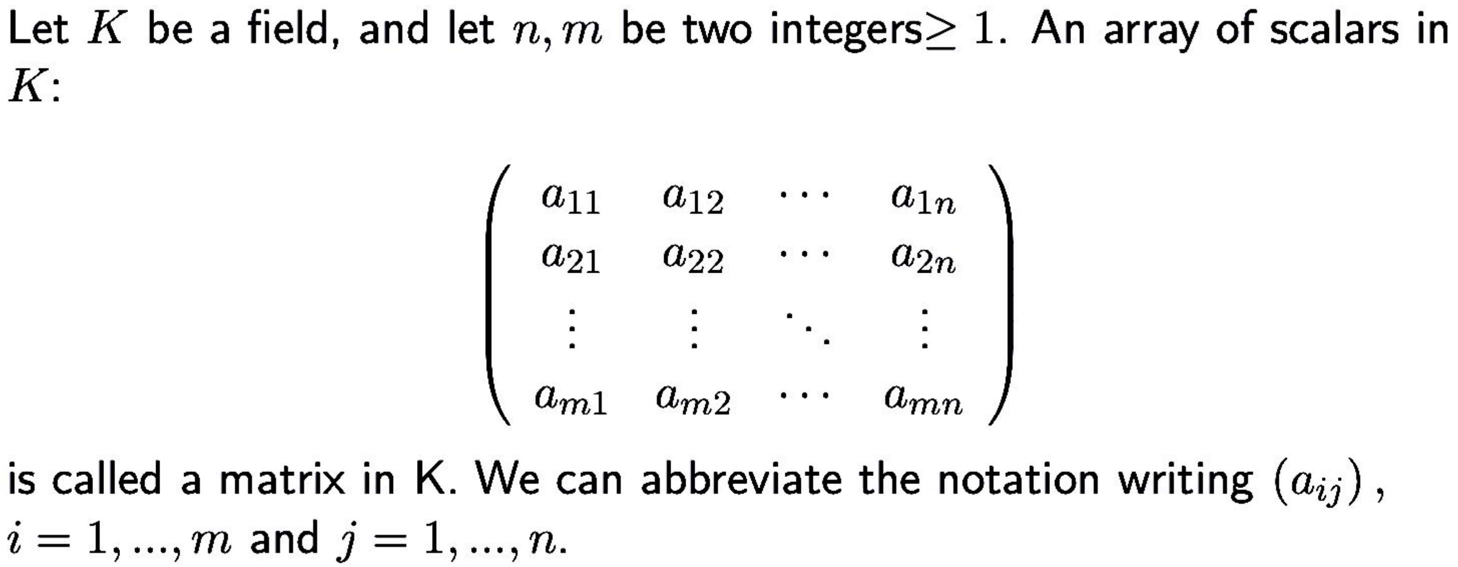
$$\begin{aligned} 3x + 3y &= 12 \\ x - 2y &= 1 \end{aligned}$$

$$\begin{pmatrix} 3 & 3 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 12 \\ 1 \end{pmatrix}$$

$$\boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} 12 \\ 1 \end{pmatrix} \text{ and } A = \begin{pmatrix} 3 & 3 \\ 1 & -2 \end{pmatrix}$$

Linear Algebra

- **MATRICES AND THEIR OPERATIONS**
  - ○ **Definition of Matrices**



Let K be a field, and let n, m be two integers≥ 1. An array of scalars in K:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

is called a matrix in K. We can abbreviate the notation writing (a_{ij}) , $i = 1, ..., m$ and $j = 1, ..., n$.

Linear Algebra

- **MATRICES AND THEIR OPERATIONS**
  - **Definition of Matrices**
    - We call a_ij the ij-entry of the matrix, and the ith row is defined as

$$A_i = (a_{i1}, a_{i2}, \ldots, a_{in})$$

    - The jth column is denoted as

$$A^j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

Linear Algebra

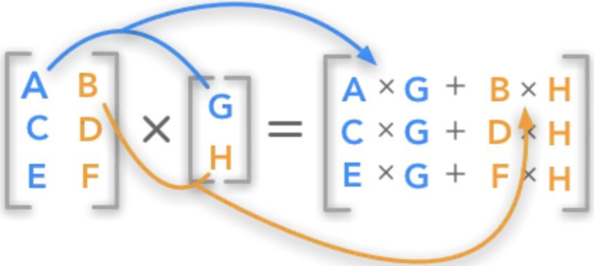
- **MATRICES AND THEIR OPERATIONS**
  - **Addition of Matrices**

$$\begin{pmatrix} 1 & -1 & 0 \\ 2 & 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 1 \\ 2 & 1 \end{pmatrix} = ?$$

  - **Matrix Multiplication**
    - Example: multiply matrices A and B

$$c_{ij} = \sum_{h=1}^{k} a_{ik} b_{kj}$$

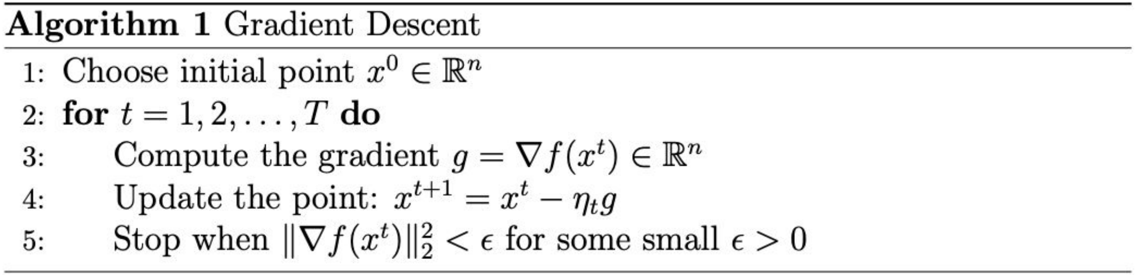
$$A = \begin{bmatrix} -4 & 5 & -3 & -2 \\ -1 & 0 & 1 & 2 \\ 3 & 4 & 6 & 7 \\ 8 & 9 & 10 & 11 \end{bmatrix}, B = \begin{bmatrix} 12 & 5 \\ 13 & 0 \\ 3 & 14 \\ 4 & 8 \end{bmatrix}$$



$$\begin{bmatrix} A & B \\ C & D \\ E & F \end{bmatrix} \times \begin{bmatrix} G \\ H \end{bmatrix} = \begin{bmatrix} A \times G + B \times H \\ C \times G + D \times H \\ E \times G + F \times H \end{bmatrix}$$

Optimizations Introduction

- **GRADIENT DESCENT**
    - Gradient Descent is an optimization algorithm and it finds out the local minima of a differentiable function.
    - It is a minimization algorithm that minimizes a given function. $\min_{x \in \mathbb{R}^d} f(x)$



Algorithm 1 Gradient Descent

```
Choose initial point $x^0 \in \mathbb{R}^n$
for $t = 1, 2, \ldots, T$ do
    Compute the gradient $g = \nabla f(x^t) \in \mathbb{R}^n$
    Update the point: $x^{t+1} = x^t - \eta_t g$
    Stop when $\|\nabla f(x^t)\|_2^2 < \epsilon$ for some small $\epsilon > 0$
```

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

- **GRADIENT DESCENT**
  - Motivation #0: **Moving to the Nearest Valley**
    - Gradient descent is a local optimization algorithm, which means that it converges to a nearby local minimum.
    - We take steps in the opposite direction $-\nabla \mathbf{f(x)}$ and gradually move towards such a local minimum.
    - Convex function
      - ❏ For a strictly convex function where the minimizer exists and is unique, gradient descent will be moving towards the same local minimum (a global minimum) regardless of where it begins.
    - Nonconvex function
      - ❏ For a nonconvex function, our choice of the initial point and step size will determine which local minimum (or saddle point) we arrive at.

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

- **GRADIENT DESCENT**
  - Motivation #0: Moving to the Nearest Valley
    - Convex function (Figure 1)
    - Nonconvex function (Figure 2)

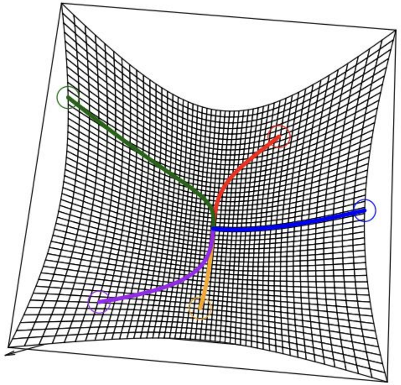


Figure 1.Gradient descent on a convex function with random initializations

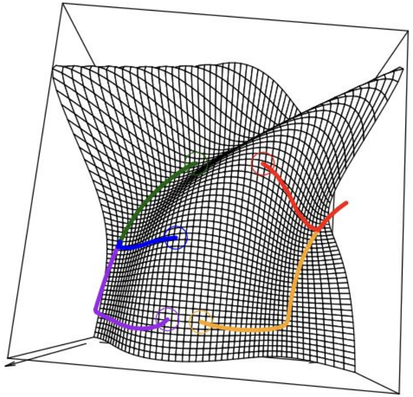


Figure 2. Gradient descent on a nonconvex function with random initializations

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

- **GRADIENT DESCENT**
  - Motivation #1: **Descent Directions**
    - There are many ways to motivate this algorithm. One is to notice that if we were at a point x and moved in a direction v with step-size $\eta > 0$

$$f(x + \eta v) \geq f(x) + \eta v^T \nabla f(x)$$

    - Goal: $v^T \nabla f(x) \leq 0$ ***Why?*** ($-\nabla$ f(x) always gives us a descent direction, otherwise we're moving to a strictly worse point)
      - ❏ Such directions (which make a larger than 90-degree angle with the gradient) are typically called descent directions (for f at x).

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

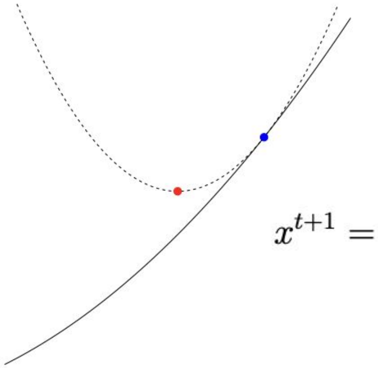
- **GRADIENT DESCENT**
  - ○ Motivation #1: **Gradient Descent as Minimizing the Local Linear Approximation**
    - A more interesting way to motivate GD (which will also be subsequently useful to motivate mirror descent, the proximal method and Newton's method) is to consider minimizing a linear approximation to our function (locally).
      - ❑ Constrained Version

$$x^{t+1} = \arg\min_{y \in \mathbb{R}^d} f(x^t) + \nabla f(x^t)^T (y - x^t)$$
$$\text{s.t. } \frac{1}{2}\|y - x^t\|_2^2 \leq \epsilon,$$

      - ❑ Unconstrained Version

$$x^{t+1} = \arg\min_{y \in \mathbb{R}^d} f(x^t) + \nabla f(x^t)^T (y - x^t) + \frac{1}{2\eta}\|y - x^t\|_2^2$$

    - Example: local minimization problem (Figure 3)



$$x^{t+1} = x^t - \eta \nabla f(x^t).$$

Figure 3. Local quadratic approximation of a function

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

- **GRADIENT DESCENT**
  - **Choosing the Step-Size**
    - Fixed Step-Size
      - ❑ Simply select a fixed step-size η and run the algorithm with that fixed step-size.
    - Exact Line-Search
      - ❑ Once we've committed to a direction (in GD this is the direction of the negative gradient), one might consider solving the following 1D optimization problem to determine the best step-size:

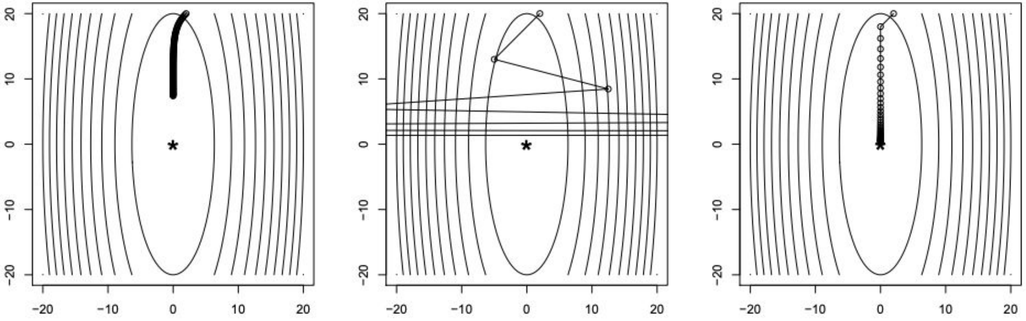
$$\eta^t = \arg\min_{\widetilde{\eta} \geq 0} f(x^t - \widetilde{\eta} \nabla f(x^t))$$

    - Backtracking Line-Search
      - ❑ The idea of backtracking line-search very roughly, is to try an aggressive (large) step-size, and reduce it by some factor if it's too big.

Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

- **GRADIENT DESCENT**
  - **Choosing the Step-Size**
    - Example
      - ❑ Examples of different steps sizes on the function $f(x) = (10x_1^2 + x_2^2)/2$. One is too small (left), one too large (middle), and one “just right” (right).



Gormley, M. (2023). Lecture 7: Gradient Descent 7.1 Gradient Descent.

Optimizations Introduction

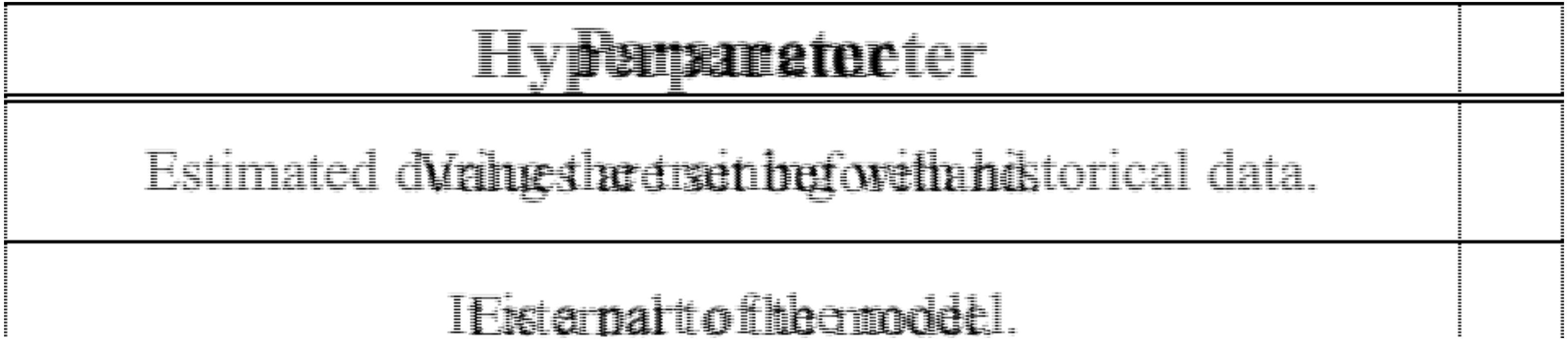
- **HYPERPARAMETERS**
  - In Machine Learning we distinguish between
    - **Model parameters:** learned by fitting model on training set
      - ❑ weights *W* and bias *b* of a layer in a Neural Network (NN)
      - ❑ parameters □$_0$ ... □*n* in a Linear Regression model $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
      - ❑ weights *W* and bias *b* of the decision hyperplane *Wx* - *b* = *0* in a Support Vector Machine (SVM)
      - ❑ Example:
        1. The coefficients (or weights) of linear and logistic regression models.
        2. Weights and biases of a nn
        3. The cluster centroids in clustering
    - **Model hyperparameters:** set by the user before training, not changed when fitting to data
      - ❑ number of hidden layers, number of neurons per layer, dropout rate, ... in a NN
      - ❑ regularization parameter **λ** for the **L^1** (Lasso) or **L^2** (Ridge) penalty term in a loss function
      - ❑ kernel type (linear, polynomial, RBF, ...) for the kernel of an SVM

Optimizations Introduction

- **HYPERPARAMETERS**
  - Here are some common examples
    - Train-test split ratio
    - Learning rate in optimization algorithms (e.g. gradient descent)
    - Choice of optimization algorithm (e.g., gradient descent, stochastic gradient descent, or Adam optimizer)
    - Choice of activation function in a neural network (nn) layer (e.g. Sigmoid, ReLU, Tanh)
    - The choice of cost or loss function the model will use
    - Number of hidden layers in a nn
    - Number of activation units in each layer
    - The drop-out rate in nn (dropout probability)
    - Number of iterations (epochs) in training a nn
    - Number of clusters in a clustering task
    - Kernel or filter size in convolutional layers
    - Pooling size
    - Batch size

Optimizations Introduction

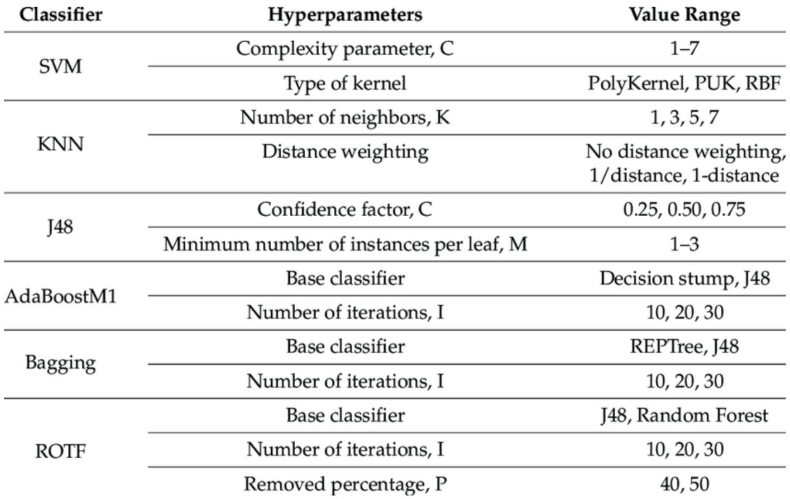
- **HYPERPARAMETERS**
  - Parameter vs. Hyperparameter



Hyperparameter Parameter	
Estimated during training with historical data. Values are set beforehand.	
Internal to the model. External to the model.	

Optimizations Introduction

- **HYPERPARAMETERS**
  - Hyperparameters and their value range of machine learning models



Classifier	Hyperparameters	Value Range
SVM	Complexity parameter, C	1–7
	Type of kernel	PolyKernel, PUK, RBF
KNN	Number of neighbors, K	1, 3, 5, 7
	Distance weighting	No distance weighting, 1/distance, 1-distance
J48	Confidence factor, C	0.25, 0.50, 0.75
	Minimum number of instances per leaf, M	1–3
AdaBoostM1	Base classifier	Decision stump, J48
	Number of iterations, I	10, 20, 30
Bagging	Base classifier	REPTree, J48
	Number of iterations, I	10, 20, 30
ROTF	Base classifier	J48, Random Forest
	Number of iterations, I	10, 20, 30
	Removed percentage, P	40, 50

Vishnu Vardhana Reddy, Karna & Elamvazuthi, Irraivan & Aziz, Azrina & Paramasivam, Sivajothi & Chua, Hui Na & Pranavanand, Satyamurthy. (2022). An Efficient Prediction System for Coronary Heart Disease Risk Using Selected Principal Components and Hyperparameter Optimization. Applied Sciences. 13. 118. 10.3390/app13010118.

Optimizations Introduction

- **HYPERPARAMETERS**
  - ○ Model performance strongly depends on hyperparameters: **how to choose them?**
    - **Basic approach:** ask "experts" of the field (= black art), try with a rule of thumb, …
    - **Our goal:** build automatic techniques to find hyperparams maximizing some metric (e.g. accuracy)
  - ○ **Categories of Hyperparameters**
    - Hyperparameter for Optimization (Hyperparameter Tuning)
      - ❑ Learning Rate
      - ❑ Batch Size
    - Hyperparameter for Specific Models
      - ❑ A number of Hidden Units
      - ❑ Number of Layers:input layers, hidden layers, and output layers

Calculus

- **CALCULUS IN MACHINE LEARNING**
    - **Gradient Descent and Backpropagation**
        - Calculus enables the optimization of machine learning models through techniques like gradient descent.
        - Backpropagation, a key part of neural network training, relies on derivatives calculated using calculus.
    - **Complex Objective Functions**
        - Calculus helps us optimize complex objective functions, crucial for model performance.
        - It allows us to handle multidimensional inputs, common in various machine learning tasks.

Calculus

- **DERIVATIVES**
  - **Derivative and Slope**
    - Slope

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

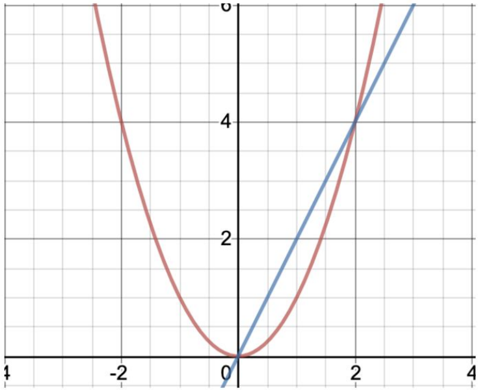
    - Example: y = 3x + 8

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{f(1) - f(0)}{1 - 0} = \frac{11 - 8}{1 - 0} = 3$$

Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf

Calculus

- **DERIVATIVES**
    - **Curves, Secant Slopes, and Derivative**
        - Given the parabola y = x^2, let’s try to find the rate of change at x = 1.



$y = x^2$ with secant $y = 2x$

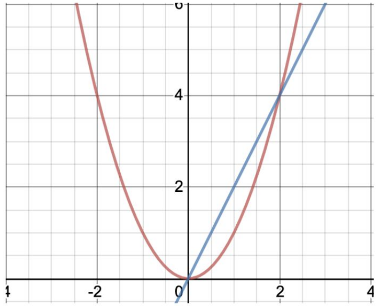
Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf

Calculus

- **DERIVATIVES**
    - **Curves, Secant Slopes, and Derivative**
        - We could approximate this value by finding the average rate of change, A, of the function between x = 1 and a close number.

$$A = \frac{f(b) - f(a)}{b - a}$$

$$\frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$



$y = x^2$ with secant $y = 2x$

Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf

Calculus

- **DERIVATIVES**
  - **Run-through of Taking the Derivative of a Function**
    - Chain rule

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$

$$\frac{d}{dx}f(g(h(x))) = f'(g(h(x)))g'(h(x))h'(x)$$

$$f'(x) = h(x)g'(x) + h'(x)g(x)$$

$$f'(x) = \frac{h(x)g'(x) - h'(x)g(x)}{h(x)^2}$$

Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf

Calculus (WHY)

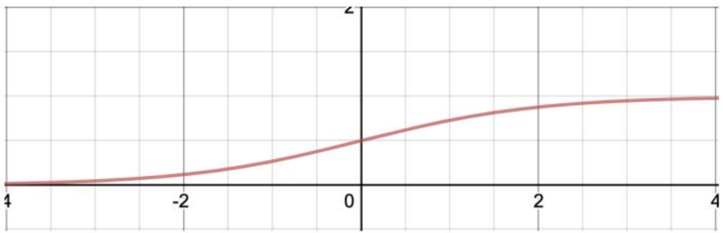
- **SIGMOID DERIVATIVE**

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

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

$$S'(x) = S(x)(1-S(x))$$

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

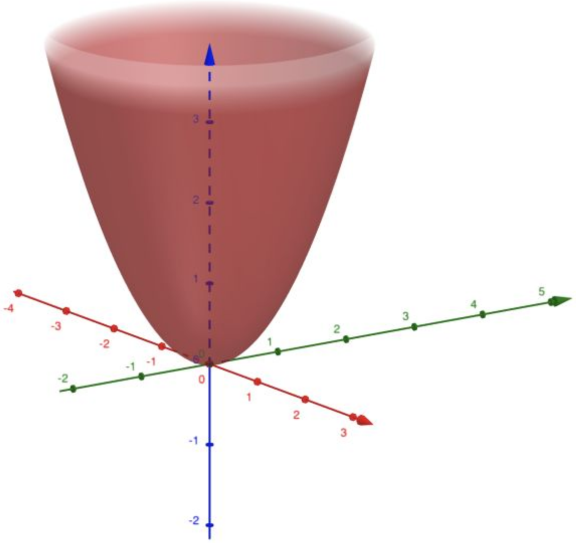


Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf

Calculus

- **MORE DIMENSIONS**
  - **Example: $z = x^2 + y^2$**
  - **Gradient**

$$\nabla z = \langle \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \rangle$$



Paraboloid $z = x^2 + y^2$

Caroline Sun (2020).– Calculus for Machine Learning – https://tjmachinelearning.com/lectures/2021/beginner/calc/calc.pdf