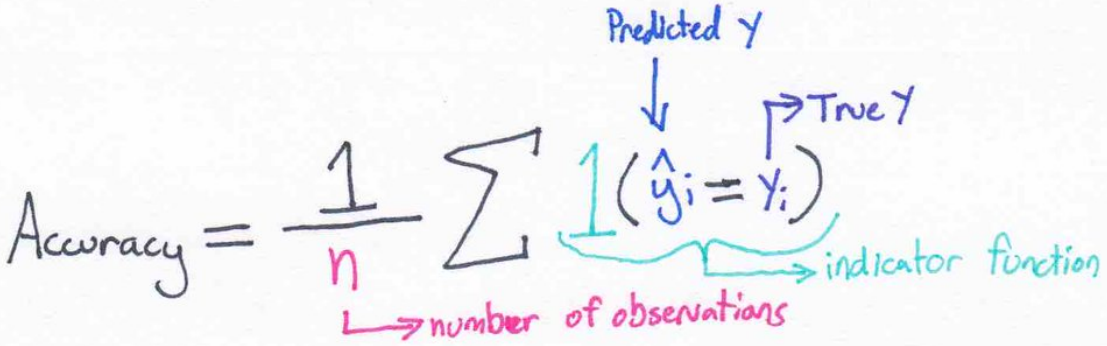

HESSIAN MATRIX
When we have a multivariate function, or multiple inputs and returning one output, the square matrix of all its second order partial derivatives is called the Hessian matrix.
LINEAR REGRESSION
PARAMETERS
FROBENIUS NORM
A measurement of the "size" of a matrix.
GENERALIZATION
How well a trained machine learning model performs on previously unseen observations
GINI INDEX
CRASCUT
1. User draws rectangle around foreground
2. Everything outside is considered background, anything inside is considered unknown
3. Gaussian measures everything inside the rectangle that looks like outside the rectangle.
GRADIENT CLIPPING
GRADIENT CLIPPING
GRADIENT DESCENT
GRADIENT ASCENT
GREEDY ALGORITHMS
GREEK LETTERS
A α alpha Δ δ delta
B β beta E ε epsilon
Γ γ gamma Z ζ zeta
EXTREMA
F1 SCORE
FPR
FEATURE IMPORTANCE
FEATURE SELECTION STRATEGIES
FEEDFORWARD NEURAL NETWORK
THE EFFECT OF ONE HOT ON FEATURE IMPORTANCE
EIGENVECTOR
ELASTIC NET
ELUs
Exponential Linear Units
ENCODING ORDINAL CATEGORICAL FEATURES
EPOCH
In artificial neural networks each time all observations have been sent through the network is called an epoch. Training neural networks typically involves multiple epochs.
ERROR TYPES
EXPLAINED SUM OF SQUARES
EXPLODING GRADIENT PROBLEM
WEIGHT DECAY
WHEN CAN WE DELETE OBSERVATIONS WITH MISSING VALUES
HIDDEN N-REPETITION
WHY IS IT CALLED A COST FUNCTION
YOUDEN'S J STATISTIC
ZERO-ONE LOSS
EARLY STOPPING ADVANTAGES
EARLY STOPPING
UNIFORM DISTRIBUTION
UNIT-STEP ACTIVATION FUNCTION
UPSAMPLING
VALIDATION CURVE
VANISHING GRADIENT PROBLEM
VIF
VARIANCE INFLATION FACTOR
VARIANCE THRESHOLDING FOR FEATURE SELECTION
VARIANCE
VECTORS
VISUALIZING RBF
WEAK LEARNERS
WINDOWING THE TEXT
TOKENLINK
TSS
Total Sum of Squares
TRAINING AND TEST ERROR
TRAINING ERROR
TPR
True Positive Rate
DROPOUT PROBABILITIES
20% of input units are dropped.
50% of hidden units are dropped.
UNDERFITTING
UNDERFLOW
TENSORS
TEST, TRAINING, VALIDATION SETS
TF-IDF
THE EFFECT OF DROPOUT ON HIDDEN UNITS
THE RANDOM IN RANDOM FOREST
THRESHOLD ACTIVATION
THRESHOLDING
SVC
SOFT-MARGIN CLASSIFICATION
SVC
RADIAL BASIS FUNCTION KERNEL
T-STATISTIC
TANH
HYPERBOLIC TANGENT ACTIVATION FUNCTION
STANDARDIZATION
STATIONARY POINTS
STEMMING WORDS
STOCHASTIC GRADIENT DESCENT
STEP WORDS

ACCURACY



Common metric in classification. Fails when in the presence of highly imbalanced classes. In those case F1 score is likely more appropriate.

BY CHRIS ALBON

AdaBoost

1. Assign every observation, x_i, an initial weight value, $w_i = \frac{1}{n}$, where n is the total number of observations.
2. Train a "weak" model. (most often a decision tree)
3. For each observation:
    - 3.1. If predicted incorrectly, w_i is increased
    - 3.2. If predicted correctly, w_i is decreased
4. Train a new weak model where observations with greater weights are given more priority.
5. Repeat steps 3 and 4 until observations perfectly predicted or a preset number of trees are trained.

Chris Albon

ADJUSTED R^2

Intuition: Once all the correct features have been added, additional features should be penalized.

$$\bar{R}^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

RSS: Residual Sum of Squares

d: Number of Features

TSS: Total Sum of Squares

n: Number of Observations

ChrisAlbon

AGGLOMERATIVE CLUSTERING

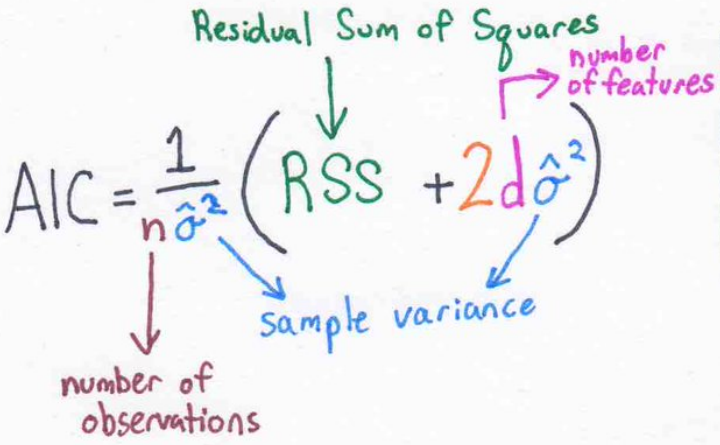
All observations start as their own cluster. Clusters meeting some criteria are merged. This process is repeated, growing clusters until some end point is reached.

ChrisAlbon

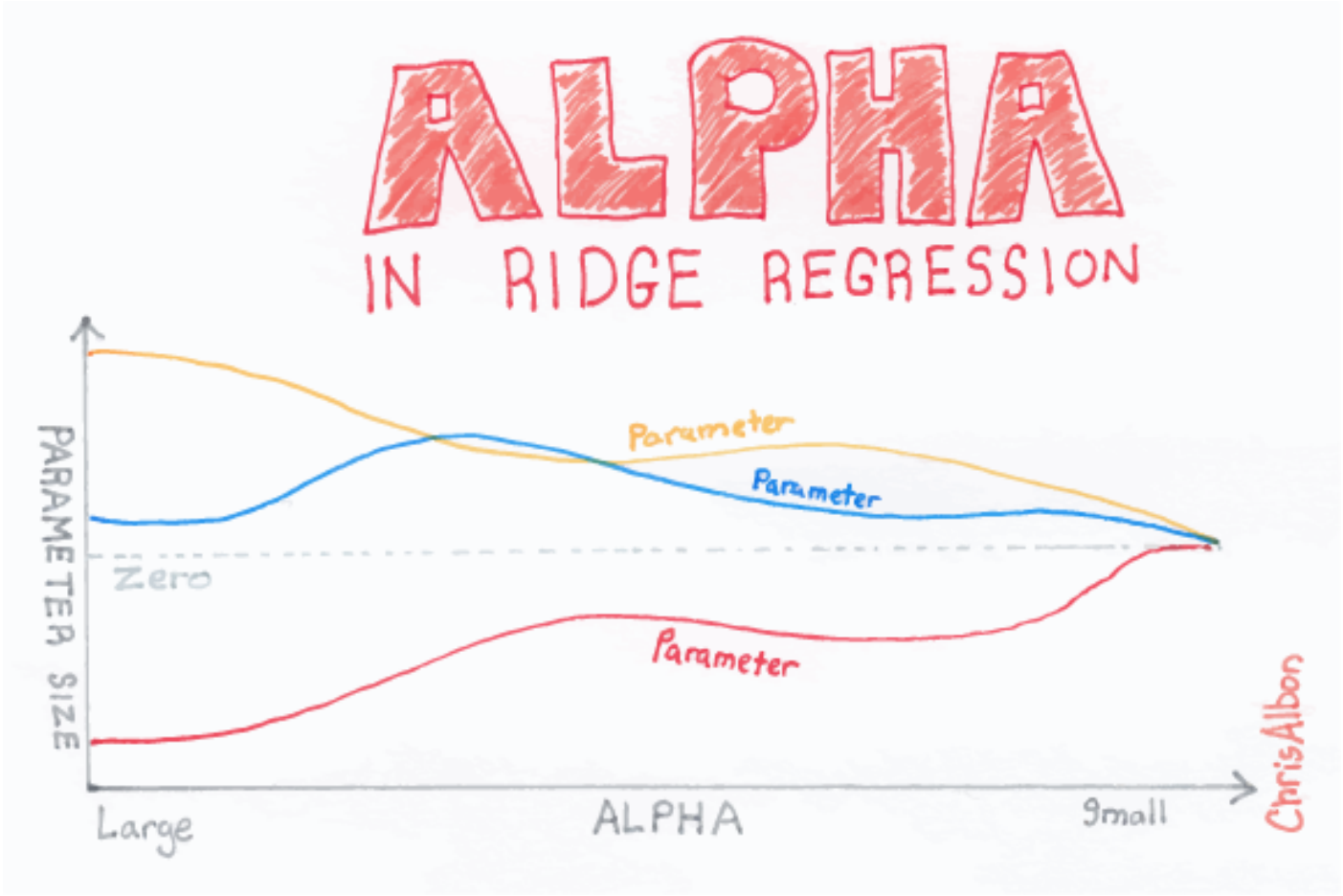
AIC

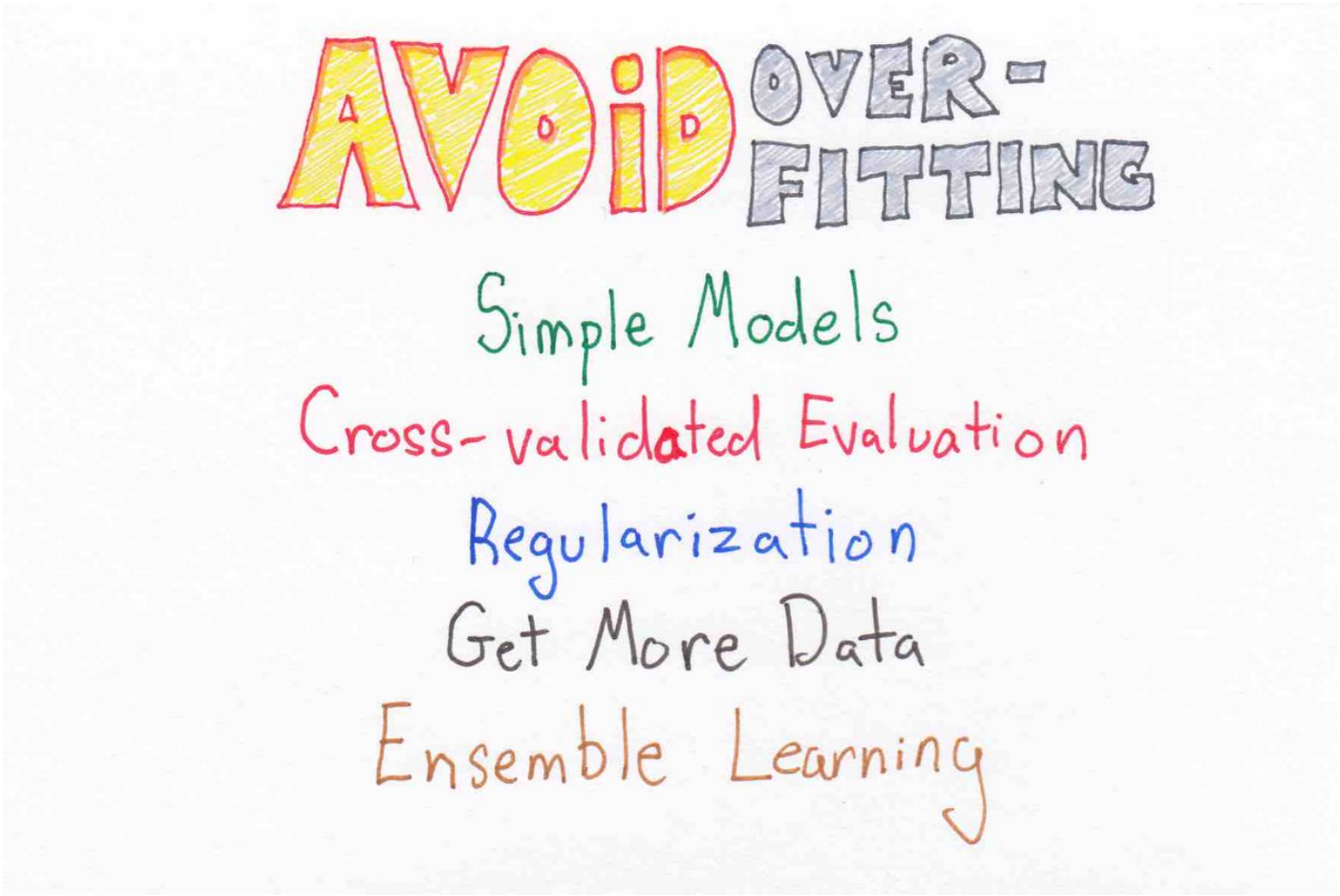
"Ey-Kye-Ih-Key"

Information Criteria



Used to compare which model is better. For example during feature selection.

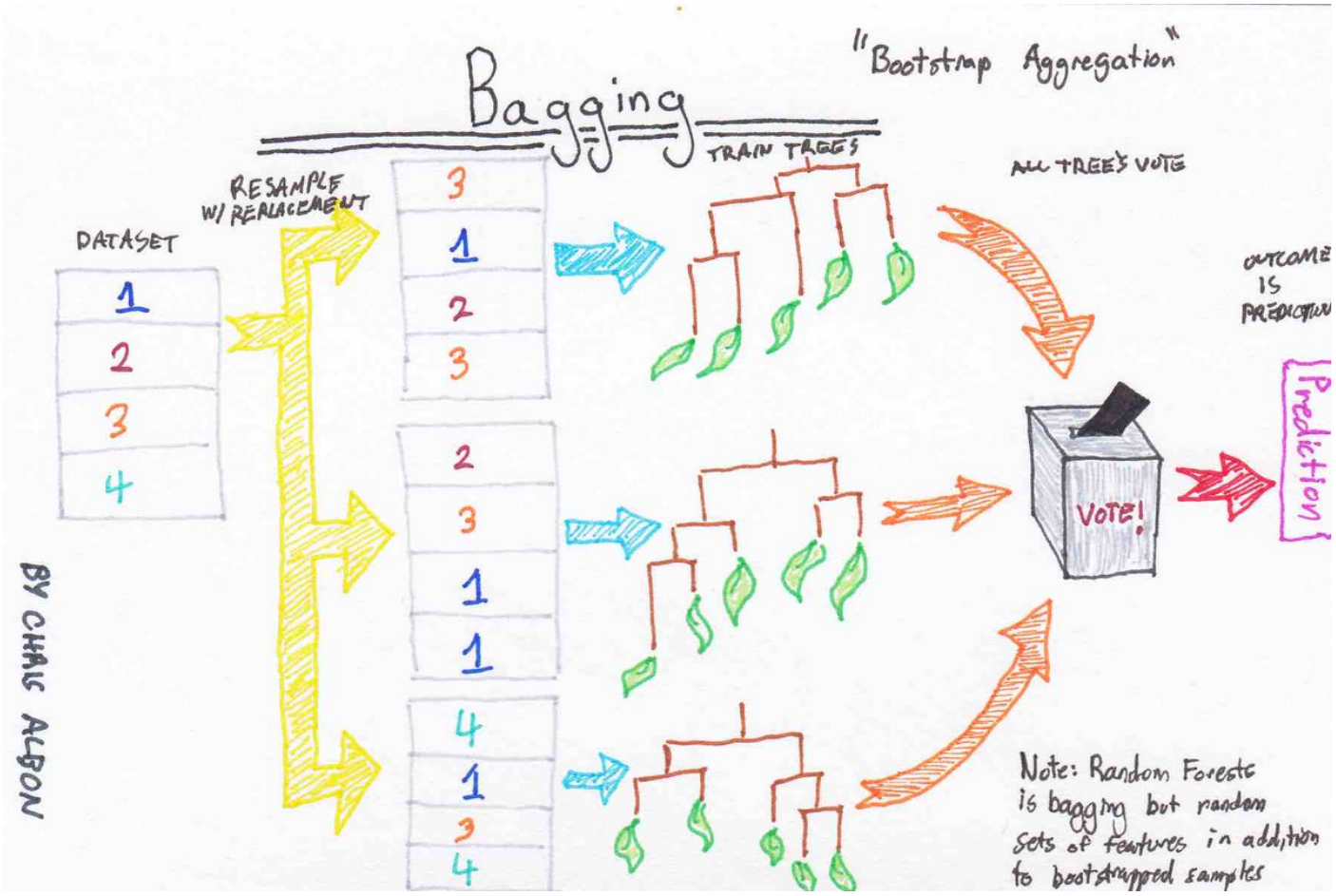

ALPHA
IN RIDGE REGRESSION
PARAMETER SIZE
Parameter
Parameter
Zero
Parameter
Large
ALPHA
Small
ChrisAlbon


AVOiD OVER-FITTING
Simple Models
Cross-validated Evaluation
Regularization
Get More Data
Ensemble Learning

BACKWARD STEPWISE SELECTION

1. Create a model, m_0, with all features, f.
2. For $K = F, K-1, \ldots, 1$:
   a) Create a model with one less feature from the last model.
   b) Repeat for all features
   c) Choose model with best evaluation metric and define as m_{K-1}
3. Select the best model from m_0 to m_f using CV.

BY CHRIS ALBON


Bagging
"Bootstrap Aggregation"
DATASET
1
2
3
4
RESAMPLE W/ REPLACEMENT
3
1
2
3
2
3
1
1
4
1
3
4
TRAIN TREES
ALL TREES VOTE
OUTCOME IS PREDICTION
VOTE!
Prediction
Note: Random Forests is bagging but random sets of features in addition to bootstrapped samples
BY CHRIS ALBON

DEEP LEARNING

BASIC PARTS

1. Data.
2. Loss function. Example: cross-entropy.
3. Optimization algorithm. Example: Adam
4. Network architecture. Example: Dense layers
5. Test data.
6. Evaluation metric. Example: Accuracy

Chris Albon

Bayes Theorem

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

POSTERIOR

LIKELIHOOD

PRIOR

MARGINAL LIKELIHOOD

Bias - Variance Tradeoff

$$Error(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\hat{f}(x) - E\left[\hat{f}(x)\right]\right]^2 + \sigma_e^2$$

$E\left[\hat{f}(x)\right]$: predicted

$f(x)$: true

$\left(E\left[\hat{f}(x)\right] - f(x)\right)^2$: $Bias^2$

How much predicted values differ from true values.

$\hat{f}(x)$: predicted

$E\left[\hat{f}(x)\right]$: average predicted value

$E\left[\hat{f}(x) - E\left[\hat{f}(x)\right]\right]^2$: Variance

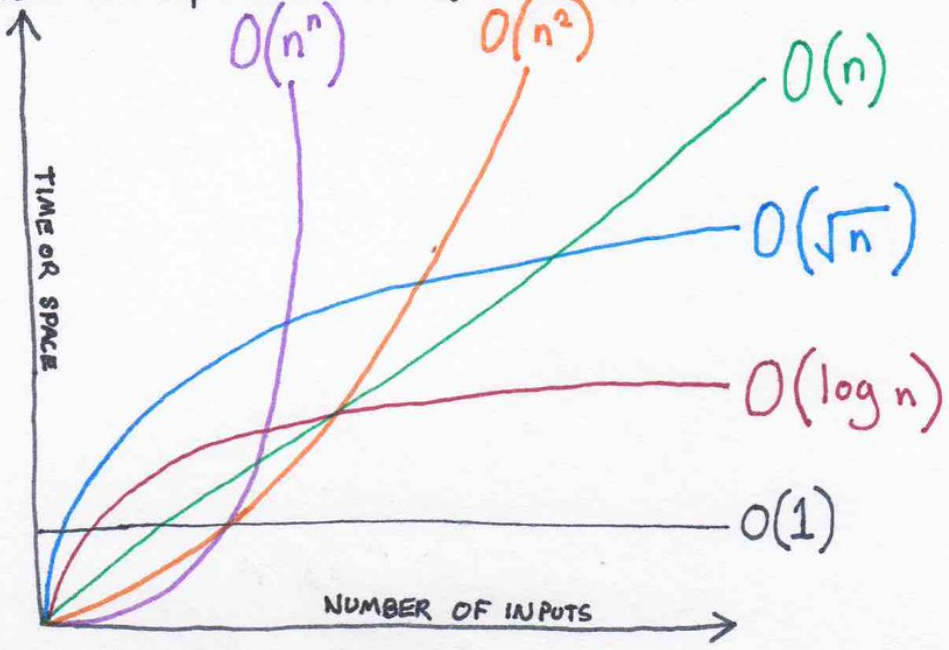
How predictions made on the same value vary on different realizations of the model

σ_e^2: irreducible error

BY CHRIS ALBON

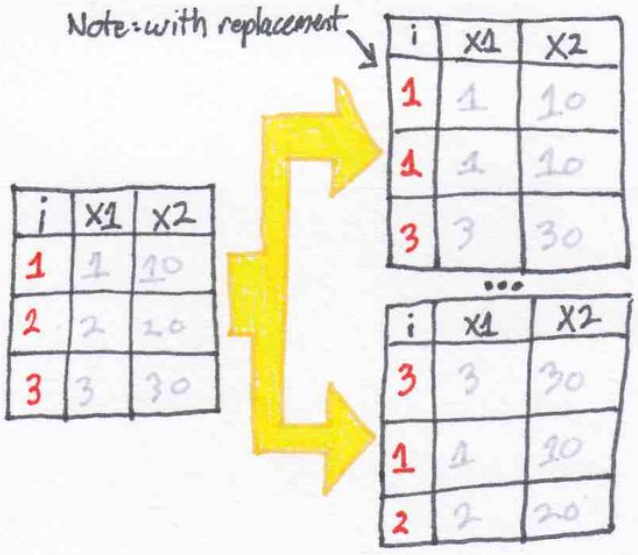
BIG O

Representation of how an algorithm's space requirement or run time increases as the number of inputs increases (in worst case scenario).



BY CHRIS ALBON

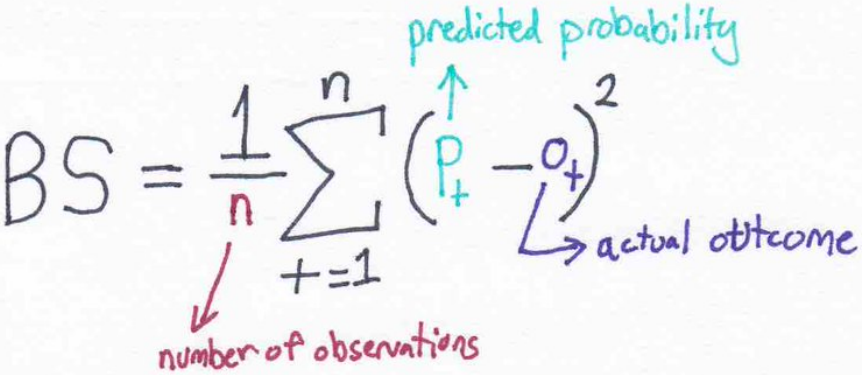
BOOTSTRAP



Bootstrap allows us to simulate obtaining many new datasets by repeated sampling with replacement from the original dataset.

Note: normally we would create many more than two bootstrapped data sets

Brier Score



Brier score shows the squared mean difference between the predicted probability of all observations with their actual outcome. The lower the Brier score, the better. Ranges between 0 and 1.

BY CHRIS ALBON

INVERSE OF REGULARIZATION STRENGTH

A hyperparameter, α (alpha), is used to control the regularization strength. For example:

$$\alpha \sum_{j=1} |\hat{\beta}_j|$$

alpha → α ; L1 norm ← $|\hat{\beta}_j|$

However, often we will see the hyperparameter C which is the inverse

$$C = \frac{1}{\alpha}$$

Chris Albon

VIF

- VIF measures the effect of collinearity among features.
- VIF measures how much the variance of a model parameter (coefficient) increases if features are correlated.
- To calculate VIF we make a feature the target of the model (instead of the dependent variable). Run the model, then calculate the R^2:

$VIF_i = \frac{1}{1 - R_i^2}$ where i is the i^{th} predictor.

$VIF_i \leq 1$ → none

$VIF_i \leq 5$ → moderate

$VIF_i \leq 10$ → high

BY CHRIS ALBON

Chi-Squared For Feature Selection

IF TWO VECTORS ARE INDEPENDENT, PROB. CATEGORY OF FEATURE IN TARGET CLASS A WILL BE THE SAME AS TARGET CLASS B.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i: count of observations in group/class i

E_i: expected number of observation in class i

We calculate between each feature and target vector. If features are independent, then feature is likely irrelevant.

Chi-Squared

$$\chi = \sum_{i=1}^{n} \left(\frac{O_i - E_i}{E_i} \right)^2$$

O_i = observed value

E_i = expected value

Tests if categorical data shows up at a rate different than random.

BY CHRIS ALBON

Coefficient of Determination, R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- y_i: true y
- $\hat{y}_i$: predicted y
- $\bar{y}$: mean y

How much of the variance in the target vector is explained by the features

Common Mathematical Notation

$\neg$	NOT
$a \in b$	a is an element of b
$\ni$	such that
$\therefore$	therefore
$\because$	because
$\Rightarrow$	then
$\Leftrightarrow$	if and only if
$\exists$	there exists
$\forall$	for all
$\parallel$	parallel
$\perp$	orthogonal, independent
$\propto$	proportional
$\sum$	summation
$\prod$	product
$\lfloor \rfloor$	floor (round down to nearest integer)
$\lceil \rceil$	ceiling (round up to nearest integer)
$X \mid Y$	x, given y
$\max()$	max value for set/list
$\underset{x}{\operatorname{argmin}} f(x)$	value of x the minimizes $f(x)$

BY CHRIS ALBON

COMMON OPTIMIZERS FOR NEURAL NETS

1. Stocastic gradient descent
2. Stocastic gradient descent with momentum.
3. RMS Prop
4. Adam

Chris Albon

COMMON OUTPUT LAYER

ACTIVATION FUNCTIONS

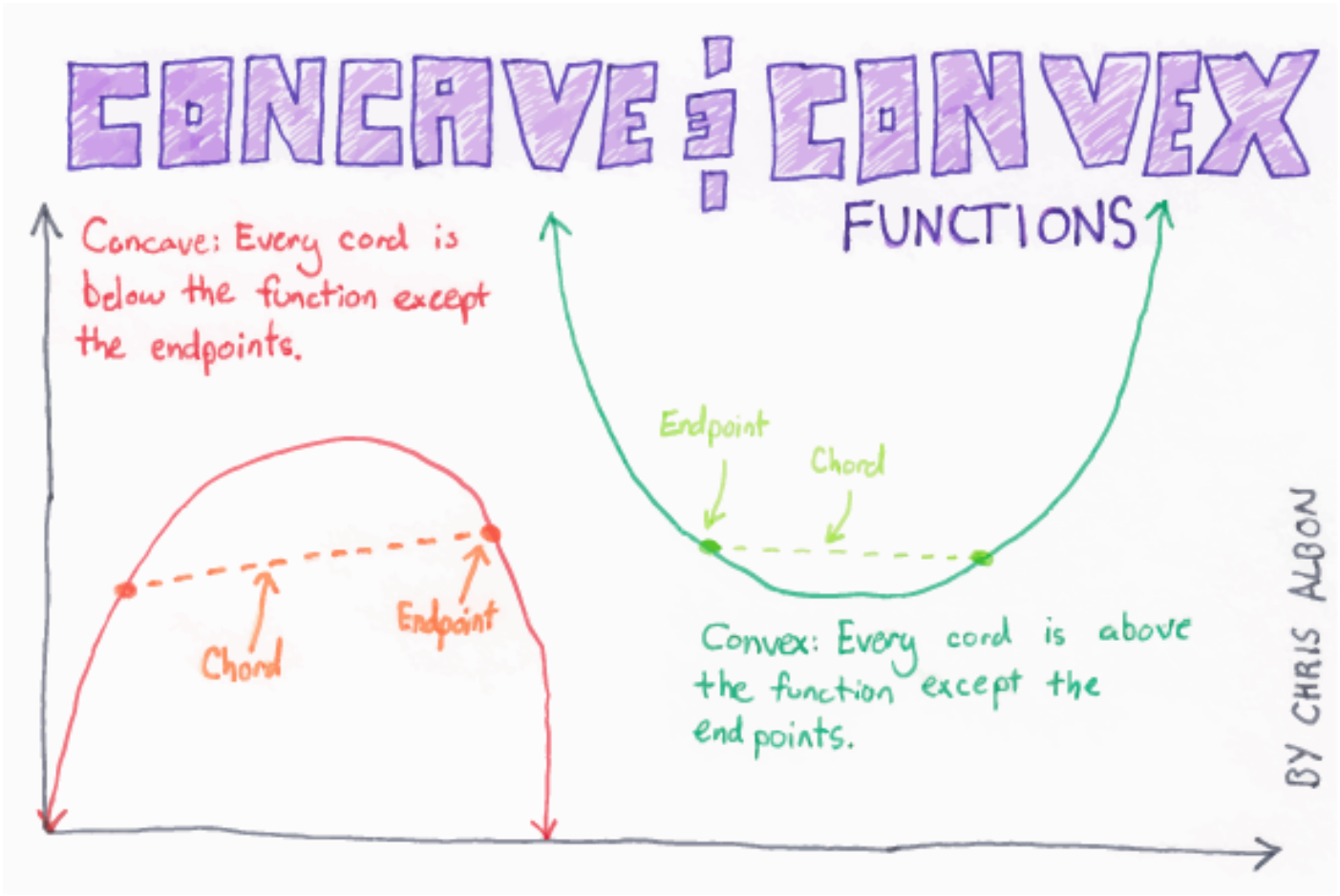
BINARY CLASSIFICATION: Sigmoid

MULTICLASS CLASSIFICATION: Softmax

REGRESSION: No activation function

Chris Albon

CONCAVE & CONVEX FUNCTIONS



Conditional Probability

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$

EXAMPLE:

$$P(A \text{ and } B) = P(\text{DRAW ACE and DRAW KING}) = \left(\frac{4}{52}\right) \times \left(\frac{4}{51}\right)$$

BY CHRIS ALBON

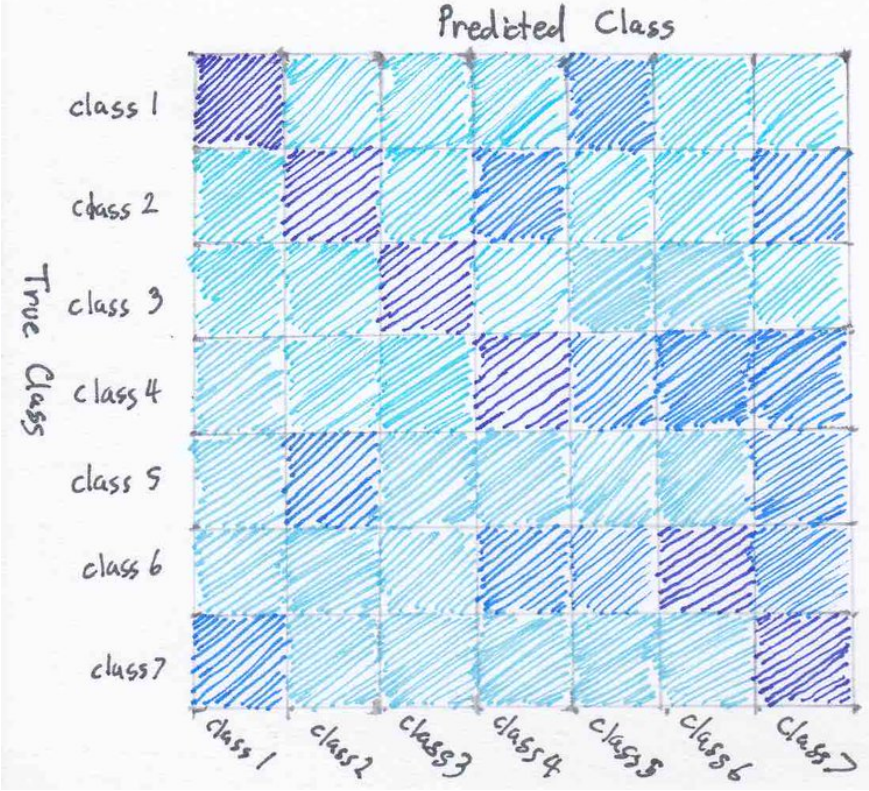
CONDITIONING

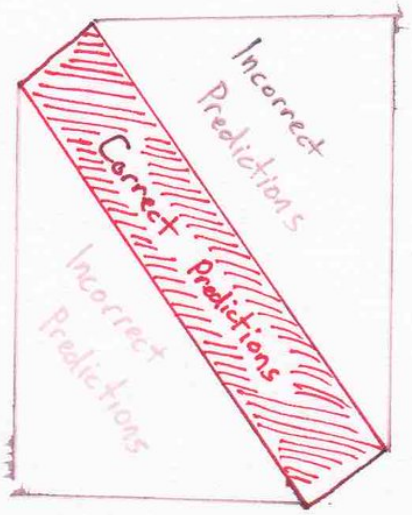
Conditioning is a measure of how much a function's outputs change when its inputs change. Poorly conditioned functions are highly sensative to rounding errors that can happen in numerical computing.

Chris Albon

Confusion Matrix

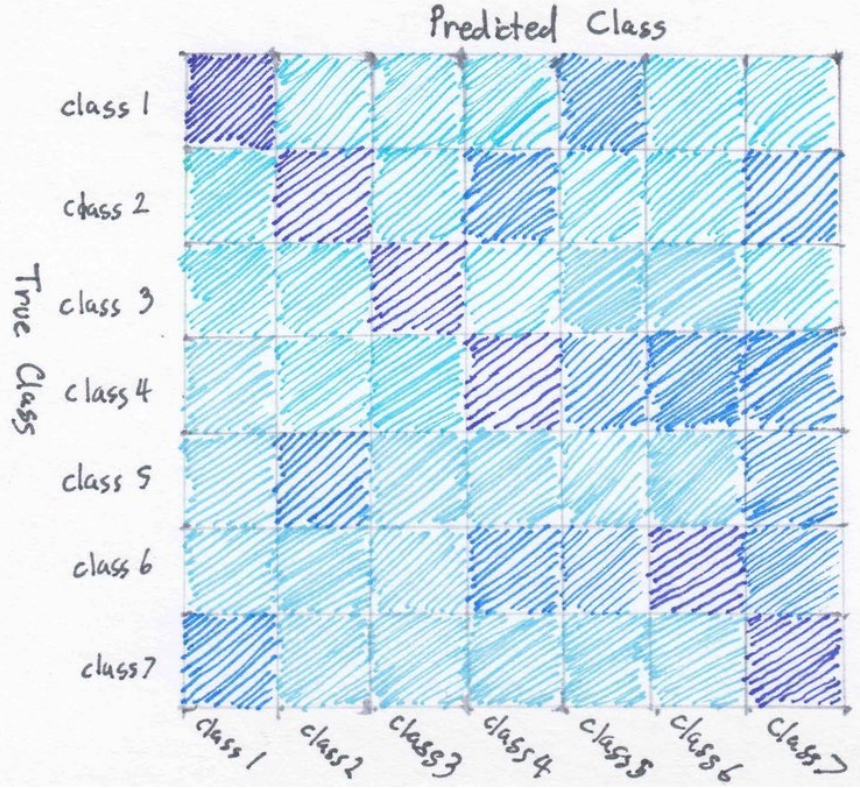
BY CHRIS ALBON

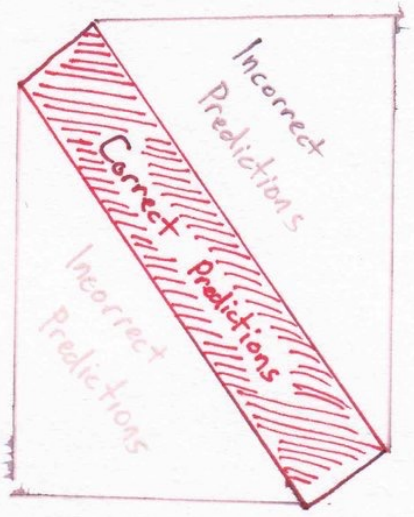




Confusion Matrices are heat maps to visualize the accuracy of a binary or multiclass classifier.

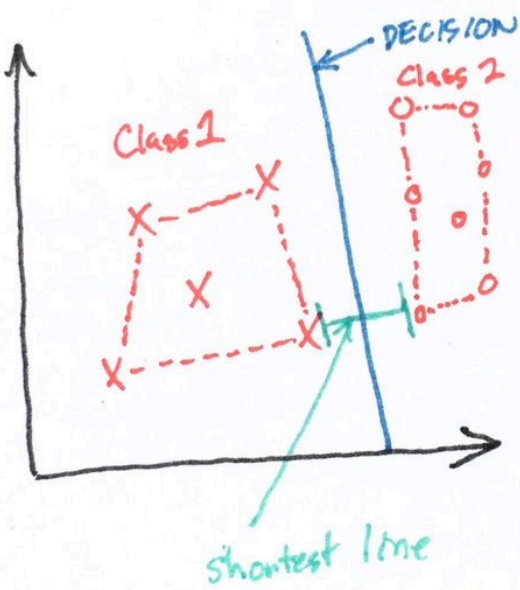
Confusion Matrix





Confusion Matrices are heat maps to visualize the accuracy of a binary or multiclass classifier.

Convex Hull's Relation To Support Vector Machines



In SVM the line of maximum margin is ~~[illegible]~~ orthogonal to the line connecting the closest two points from each group's convex hull.

Correlation of Error Terms

If there is correlation in the error terms, we will have underestimated STANDARD ERRORS.

HETEROSCEDASTICITY

Breaks Gauss-Markov Theorem:

NOT BLUE. CAUSES ESTIMATES OF OLS VARIANCE TO BE BIASED.

Correlation

I'm drunk

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}.)}{\sqrt{\sum_{i=1}^{} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{} (y_i - y_2)^2}}$$

Cost functions, loss functions, and error functions mean the same thing. They are the objective function we are trying to train a model that minimizes or maximizes, depending.

BY CHRIS ALBON

Cost functions, loss functions, and error functions mean the same thing. They are the objective function we are trying to train a model that minimizes or maximizes, depending.

Covariance and Correlation

Correlation

$$\frac{E(X_i - \bar{X})(Y_i - \bar{Y})}{(\sigma_x \ \sigma_y)}$$

$E(X_i - \bar{X})(Y_i - \bar{Y})$ ← covariance

C_p

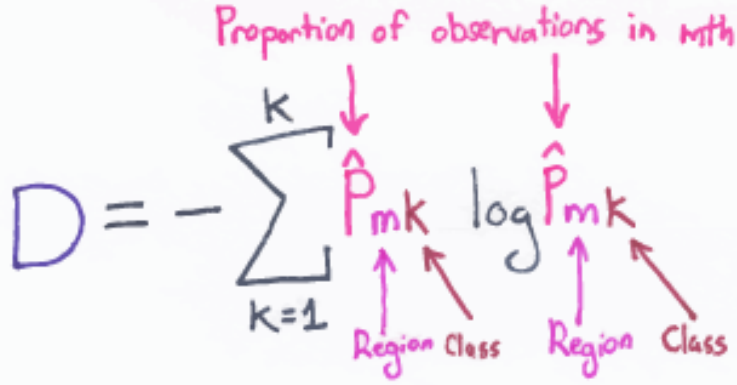
$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$$

- RSS: residual sum of squares
- d: number of features
- $\hat{\sigma}^2$: estimated error variance

$2d\hat{\sigma}^2$ is a penalty to adjust for the fact that the training data underestimates test error.

C_p is used in model selection to compare different models.

CROSS-ENTROPY



The closer D is to 0 the purer the classes.

$0 \leq -\hat{P}_{mk} \log \hat{P}_{mk}$, so the larger the value the less pure.

BY CHRIS ALBON

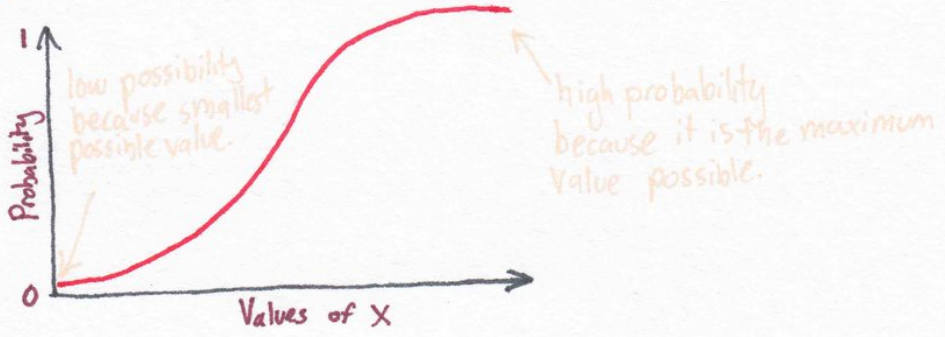
Cumulative Distribution Function

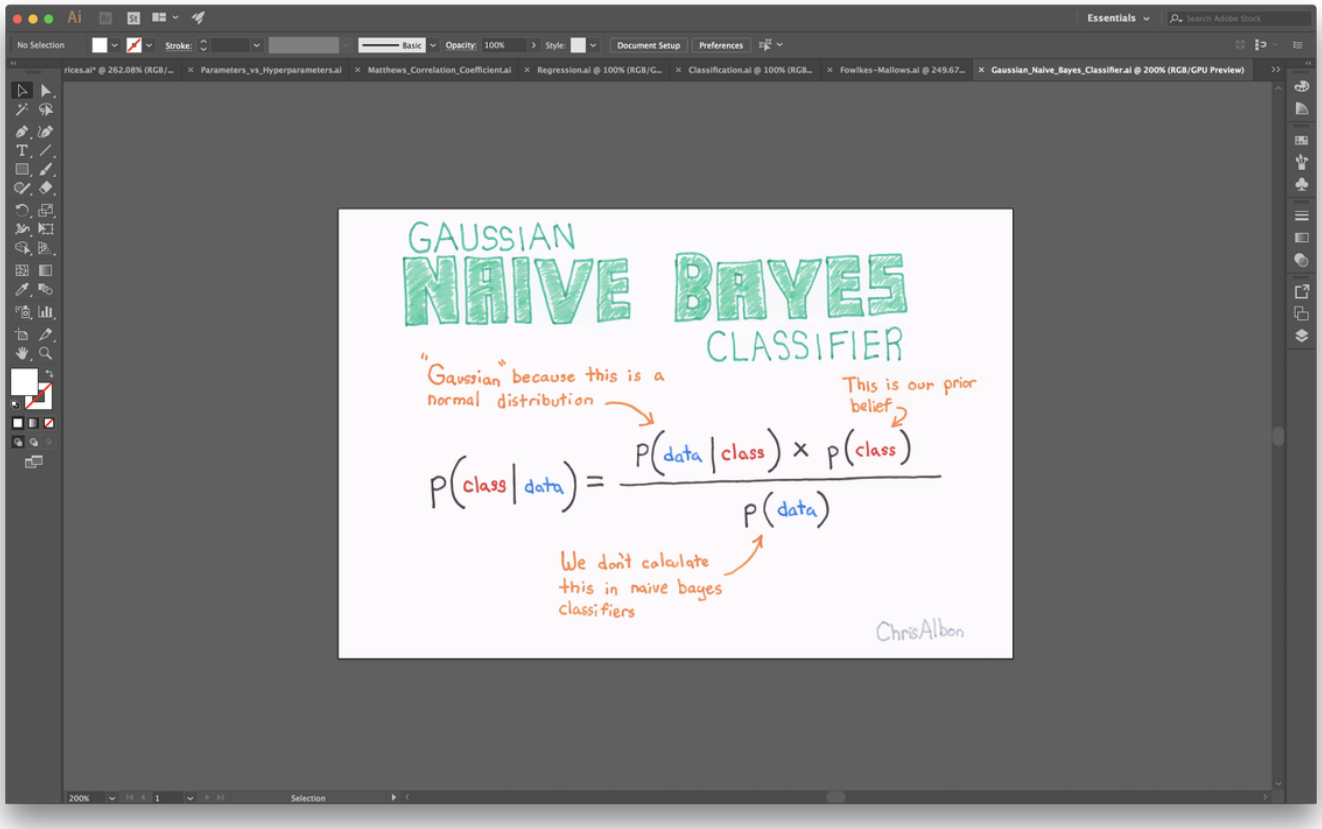
$$f(x) = Pr[X \leq x]$$

where $x \in X$

CDF tells us the probability a random variable/function returns a value less than some specified value.

CDF is the accumulation of the ~~possible~~ probabilty of values up to some value.




GAUSSIAN
NAIVE BAYES
CLASSIFIER
"Gaussian" because this is a normal distribution
This is our prior belief
P(class | data) = P(data | class) × P(class) / P(data)
We don't calculate this in naive bayes classifiers
ChrisAlbon

Curse of DIMENSIONALITY

As the dimensionality of the features space increases, the number Configurations can grow exponentially, and thus the number of configurations covered by an observation decreases.

Chris Albon

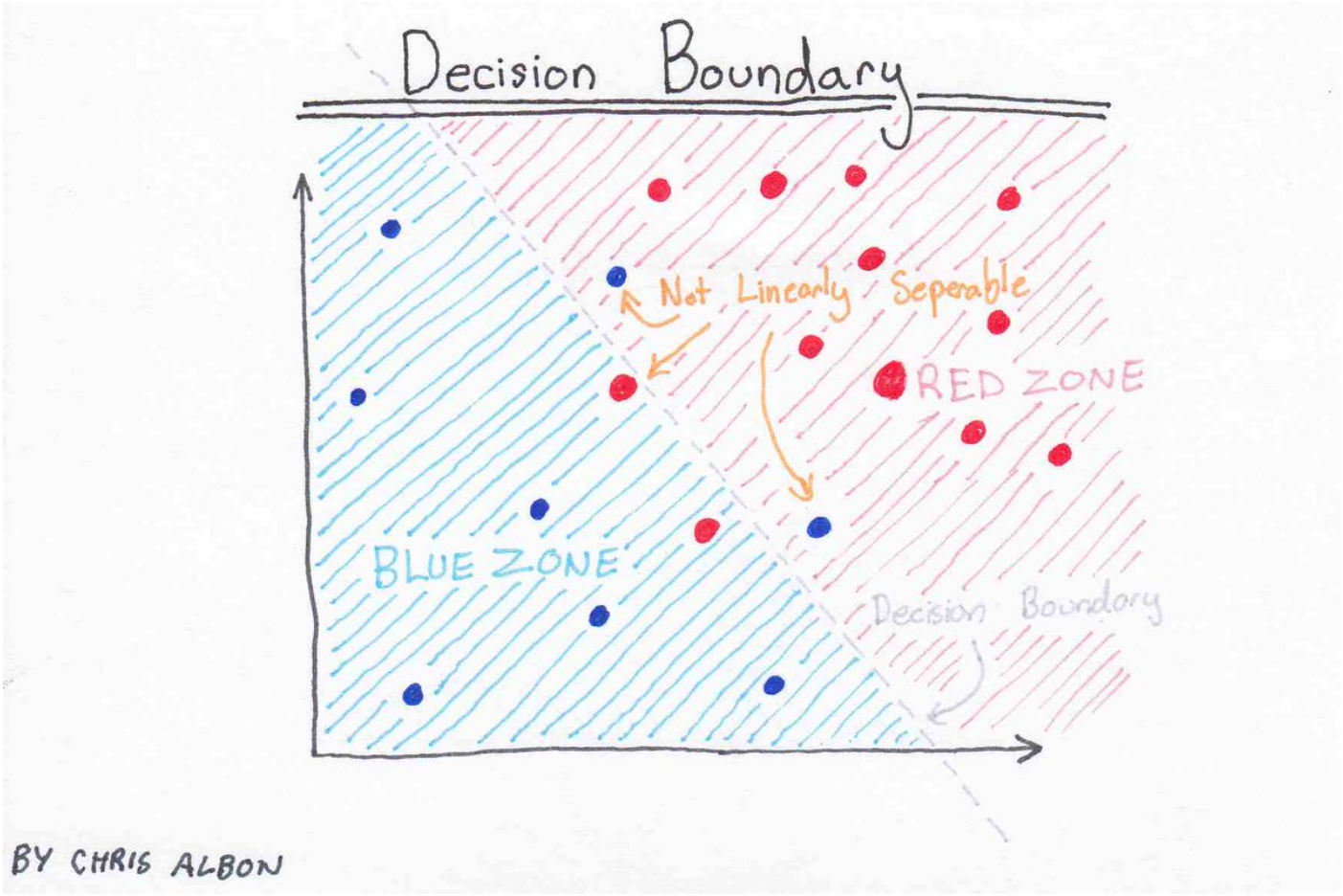
DATA-GENERATING DISTRIBUTION

The unseen distribution of the population.

DATASET AUGMENTATION

- Often used in image recognition.
- Creates noise by rotating, scaling, shifting etc... images in computervision problems. Other ways of injecting noise are used in other problems.
- Can greatly reduce generalization error.

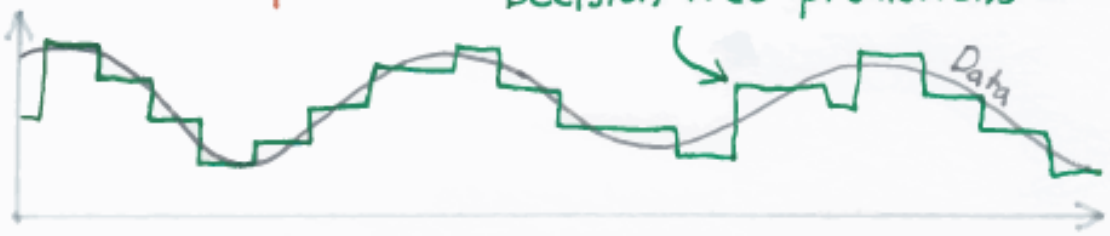
Chris Albon


Decision Boundary
Not Linearly Seperable
RED ZONE
BLUE ZONE
Decision Boundary
BY CHRIS ALBON

DECISION TREE

REGRESSION

Similar to decision tree classification, however uses Mean Squared Error or similar metrics instead of cross-entropy or Gini impurity to determine splits.

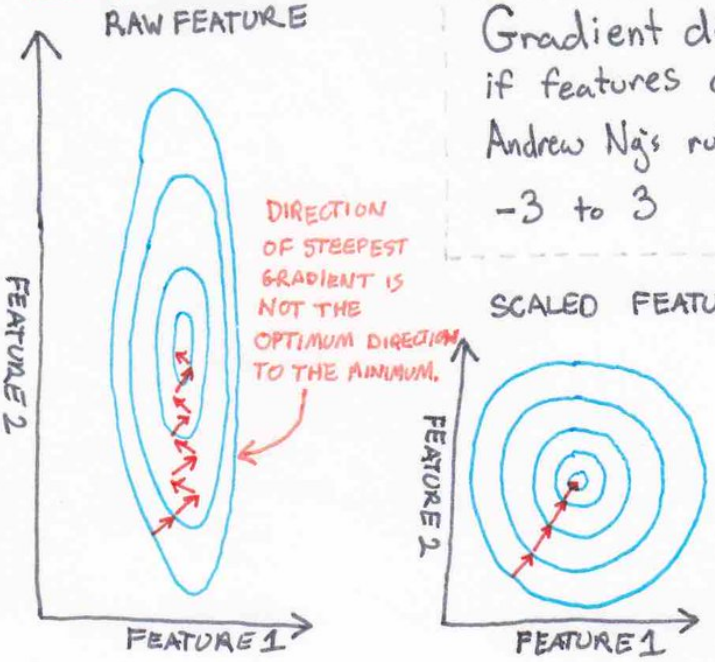


Downsampling

A strategy to handle imbalanced classes by creating a random subset of the majority of equal size to the minority class.

ChrisAlbon

EFFECT OF FEATURE SCALING ON GRADIENT DESCENT



Gradient descent will take longer if features are not similarly scaled. Andrew Ng's rule is that if x_i ranges from -3 to 3 or $-\frac{1}{3}$ to $\frac{1}{3}$ then it is fine.

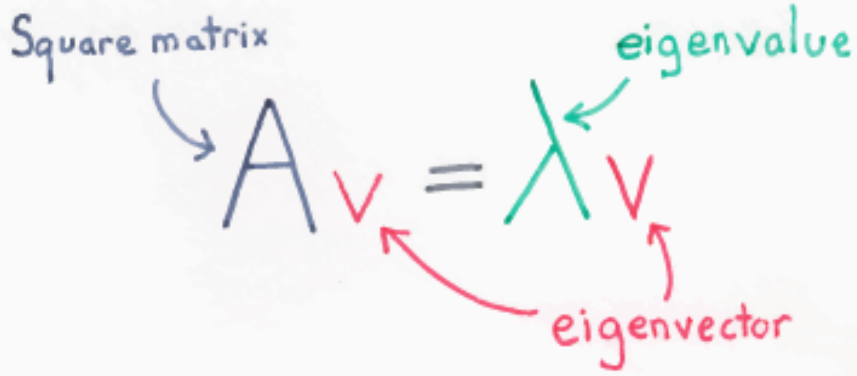
BY CHRIS ALBON

Effect of K's value in K-nearest neighbors

Small K = Low Bias, High Variance

Large K = High Bias, Low Variance

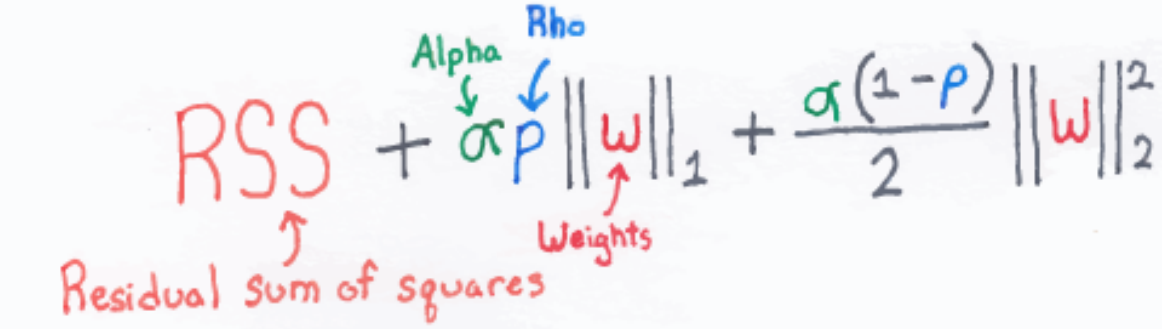
EIGENVECTOR



Chris Albon

ELASTICNET

A linear regression model that combines the L1 and L2 regularizers.



Alpha determines the regularization strength.
Rho determines the balance between L1 and L2.

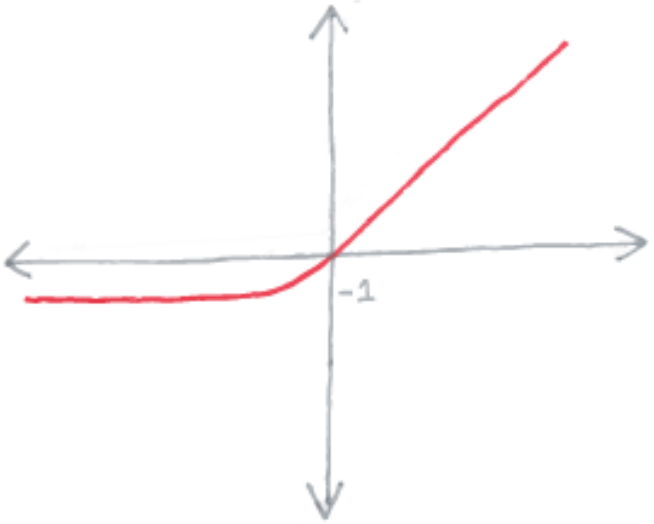
ChrisAlbon

ELUs

Exponential Linear Units

$$\phi(z) = \begin{cases} z & \text{if } z >= 0 \\ a[\exp(z) - 1] & \text{otherwise} \end{cases}$$

positive hyperparameter



Chris Albon


Error Types
Type
FALSE Positive
Type
FALSE Negative

Euclidean Distance

For vectors, p and q.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Euclidean Norm

Often used in rescaling via normalizing observation values.

$$\| x \|_2 = \sqrt{x_1^2 + x_2^2 + \ldots\ldots + x_n^2}$$

Also called: L2 norm

EXPLAINED

SUM OF SQUARES

ESS measures the amount of variance (information) in the model.

$$ESS = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

$\hat{y}_i$: Predicted value

$\bar{y}$: mean value

BY CHRIS ALBON

EXPLODING GRADIENT PROBLEM

When the gradients of the loss function with respect to the parameters in the early layers are very large. Causes unstable learning and poor performance. One potential fix is gradient clipping.

Chris Albon

F-STATISTIC

Total Sum Of Squares

Residual Sum Of Squares

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

p ← Number of Features

n ↑ Number of observations

Used by the F-test to test if groups of features are jointly statistically significant.

Chris Albon

F1 Score

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 score can be interpreted as the harmonic mean of precision and recall. Values range from 0 (bad) to 1 (good).

BY CHRIS ALBON

F1 Score

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 score can be interpreted as the harmonic mean of precision and recall. Values range from 0 (bad) to 1 (good).

FPR

False Positive Rate

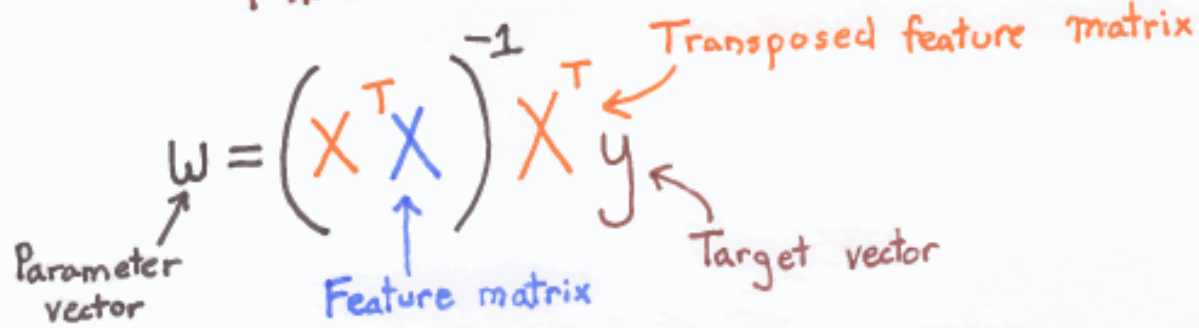
$$FPR = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

Feature Selection Strategies

1. Remove highly correlated variables
2. Run OLS and select significant features
3. Forward Selection and backwards selection.
4. Random Forest feature importance
5. Lasso

LINEAR REGRESSION

FINDING PARAMETERS



Analytical solutions are most common. However, the cost of computing pseudo-inverse of $X^T X$ means that gradient descent can be better in large data.

ChrisAlbon

FOWLKES-MALLOWS SCORE

Evaluate clusters when ground-truth is avaliable. Used to evaluate pairs of clusters. Higher score is better. Ranges between 0 and 1

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

Fowlkes-Mallows Index

True Positive

False Positive

False Negative

TP: # of pairs of observations are part of the same cluster in both y and $\hat{y}$.

FP: # of pairs of observations that are part of the same cluster in y but part of different clusters in $\hat{y}$.

FN: # of pairs of observations that are not part of the same cluster in y but part of the same cluster in $\hat{y}$.

BY CHRIS ALBON

FOWLKES-MALLOWS SCORE

Evaluate clusters when ground-truth is avaliable. Used to evaluate pairs of clusters. Higher score is better. Ranges between 0 and 1

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

Fowlkes-Mallows Index

True Positive

False Positive

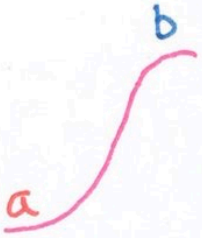
False Negative

TP: # of pairs of observations are part of the same cluster in both y and $\hat{y}$.

FP: # of pairs of observations that are part of the same cluster in y but part of different clusters in $\hat{y}$.

FN: # of pairs of observations that are not part of the same cluster in y but part of the same cluster in $\hat{y}$

Fundamental Theorem Of Calculus



$$= \text{anti-derivative}(b) - \text{anti-derivative}(a)$$

$$\boxed{F(b) - F(a)}$$

General Additive Model

$$GAM = \hat{y}_i = \beta_0 + f(x_{i1}) + f(x_{i2}) \ldots \varepsilon_i$$

- Every IV is a function.
- No interaction terms are possible.

GINI INDEX

Proportion of observations in the m^{th} leaf of K^{th} class.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right)$$

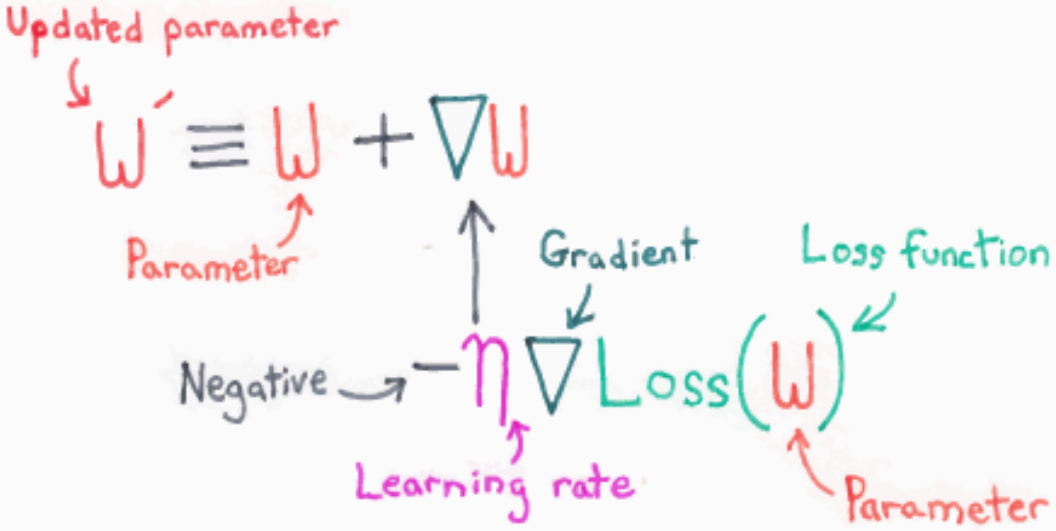
leaf Class leaf Class

Used at each node to decide which feature is best to split on.

The smaller the value of G the more purity there is in the node.

Measure of purity in tree based methods

GRADIENT DESCENT

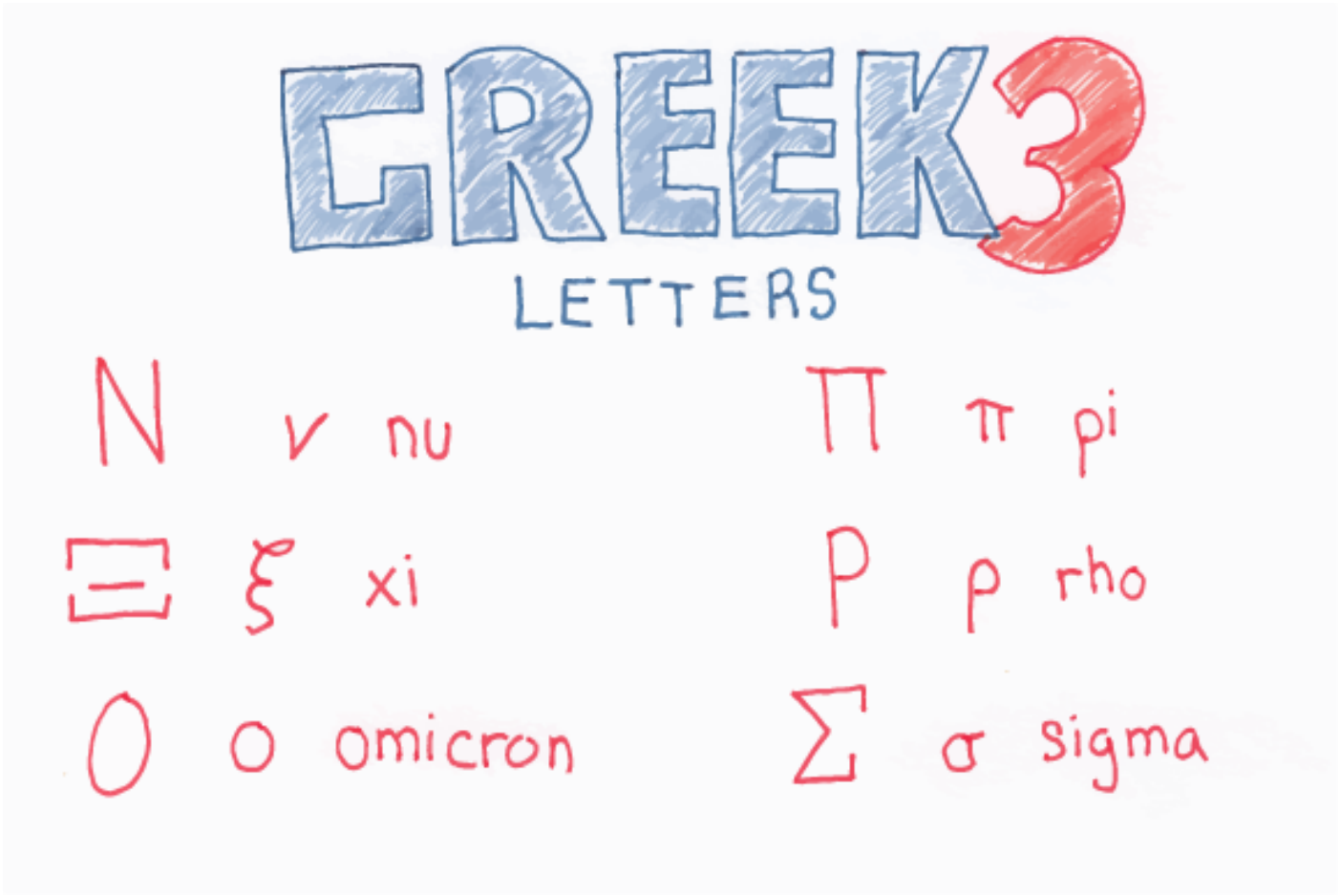


Chris Albon

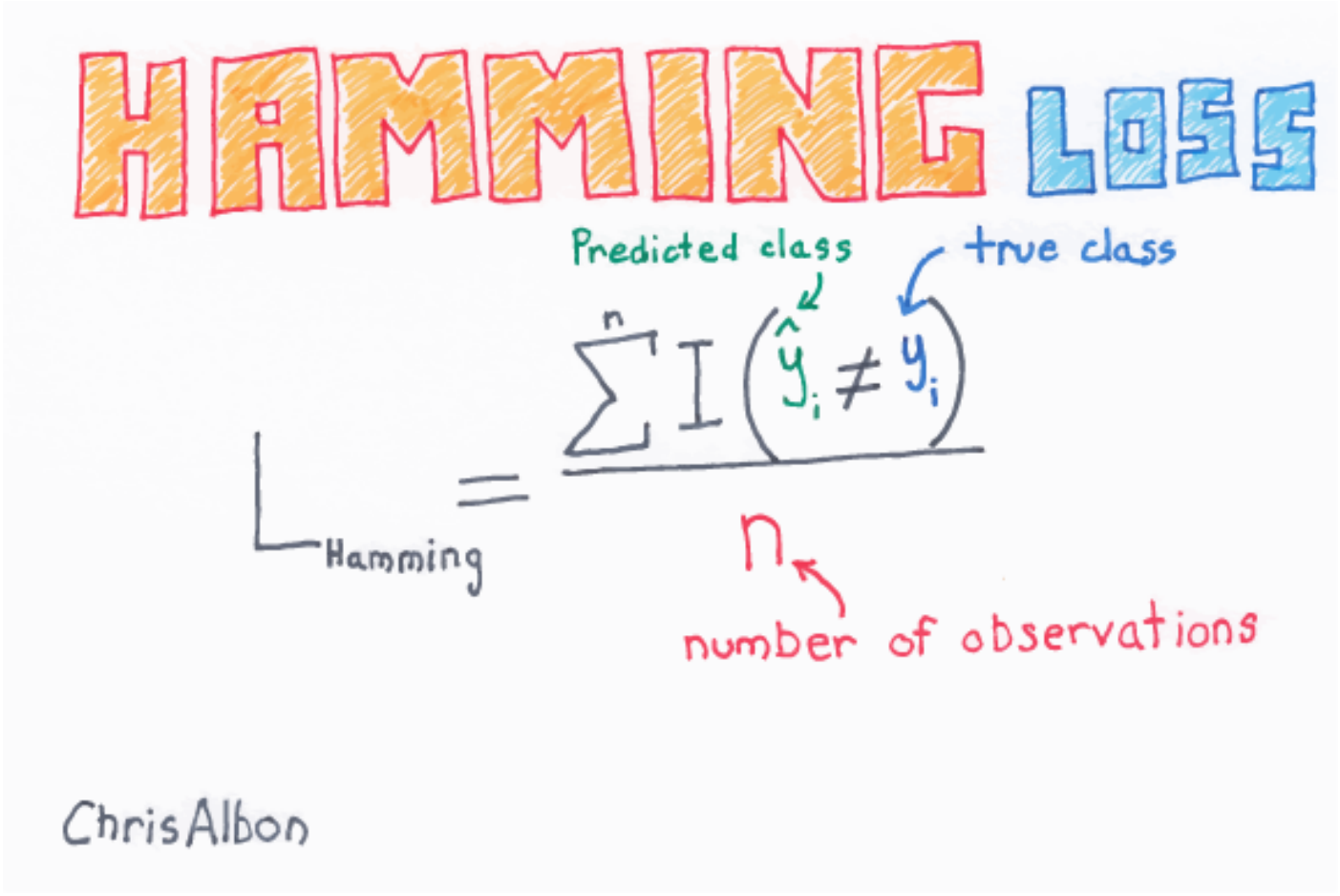
GREEK LETTERS

A	α	alpha	Δ	δ	delta
B	β	beta	E	ϵ	epsilon
Γ	γ	gamma	Z	ζ	zeta

Chris Albon


GREEK 3
LETTERS
N ν nu
Ξ ξ xi
O o omicron
Π π pi
Ρ ρ rho
Σ σ sigma

HAMMING LOSS



Chris Albon

HANDLING IMBALANCED CLASSES IN SUPPORT VECTOR MACHINES

In SVMs, the hyperparameter C determines the penalty for misclassification. To handle imbalanced classes we can weight C by class:

$$C_K = C * w_K$$

C_K: penalty for class K

w_K: weight inversely proportional to class K's frequency

The goal is to prevent the majority class from overwhelming minority class.

Chris Albon

HANDLING OUTLIERS

1. If due to an error: drop, mark as missing value, mark as possible error.
2. If a legitmate but extreme value: decide if it is genuinely a member of the population we are try to address with our model.

Chris Albon

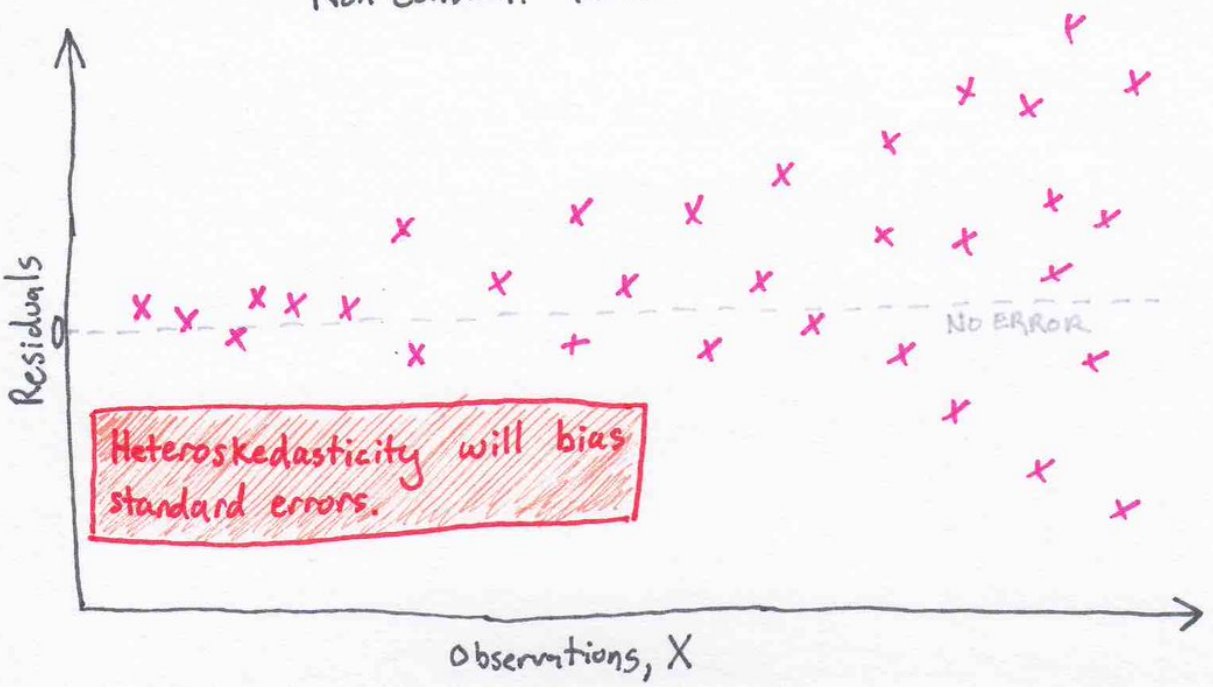
HESSIAN MATRIX

When we have a function taking in multiple inputs and returning one output, the square matrix of all its second order partial derivatives is called the Hessian matrix.

Chris Albon

HETEROSKEDASTICITY

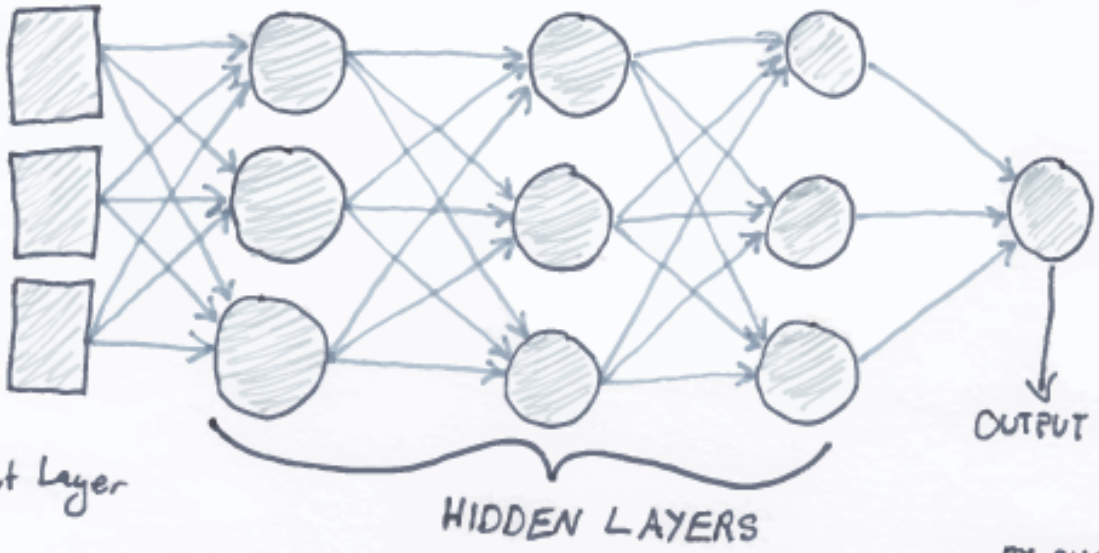
Non-constant Variance in Errors.



BY CHRIS ALBON

HIDDEN LAYER

CALLED HIDDEN LAYERS BECAUSE VALUES ARE NOT IN THE DATA. RATHER, VALUES ARE DETERMINED BY WHAT IS USEFUL FOR MODELLING RELATIONSHIP.



BY CHRIS ALDON

HINGE LOSS

Class labels are +1 and -1

$$L_{\text{Hinge}}(y, w) = \max\{1 - wy, 0\}$$

True y — Raw output of classifier — True y

Used in support vector classifiers. When y and w have the same sign and $|w| \geq 1$ then Hinge Loss is 0. Otherwise, Hinge Loss increases as w increases.

BY CHRIS ALBON

HOW NORM PENALTIES WORK

- L1 and L2 norm penalties shrink parameters toward zero.
- The benefit comes from less variance in parameter values, not necessarily small values.
- Zero is typically used because it is neither positive or negative.

Chris Albon

How to avoid overfitting

Simple models

Cross-validation

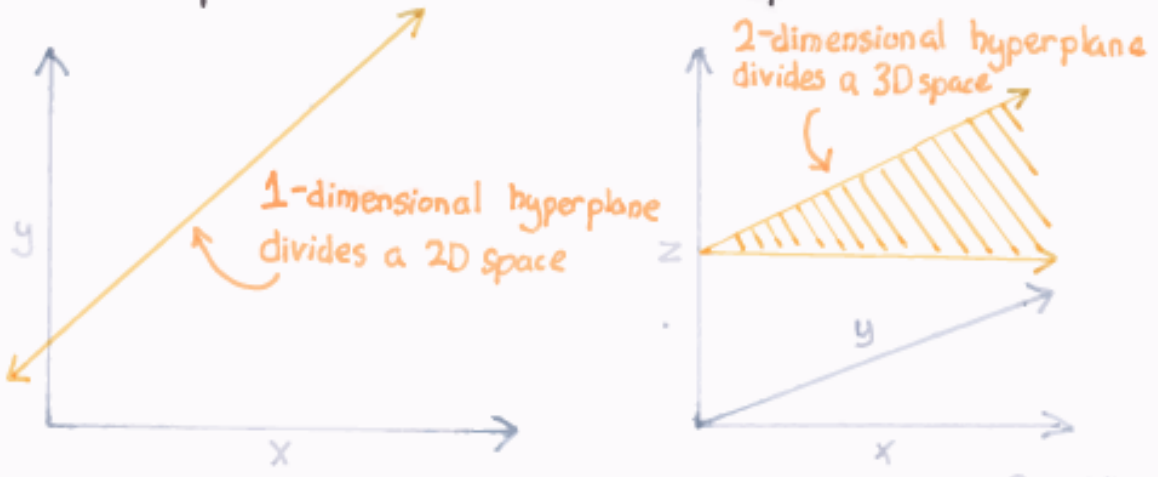
Regularization

Get More Data

Ensemble Models

HYPERPLANE

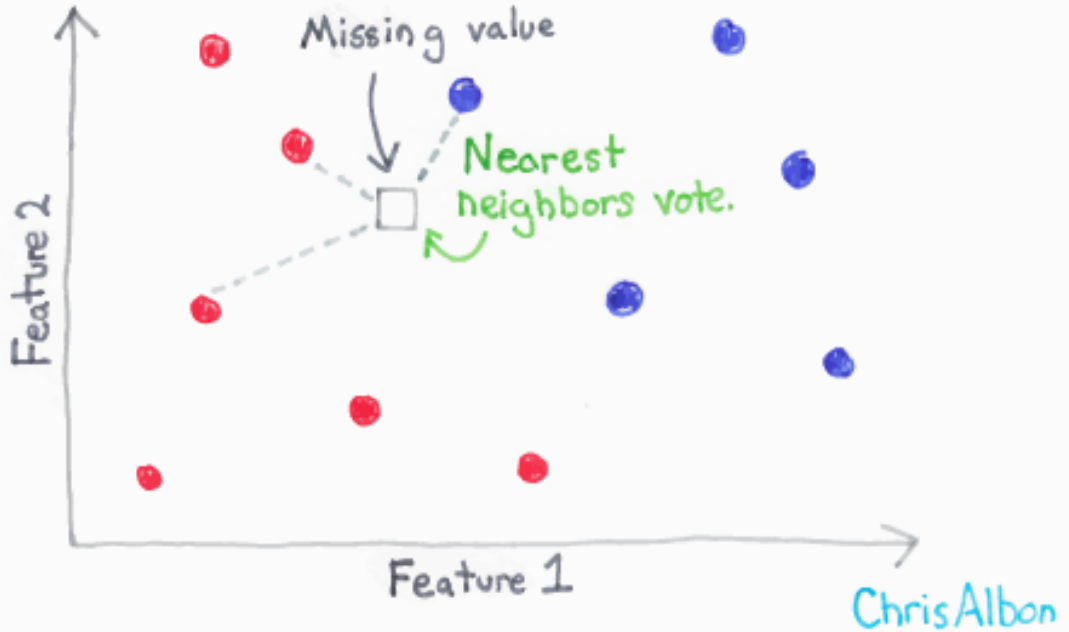
In an n-dimensional space, a hyperplane is an n-1 plane that divides that space.



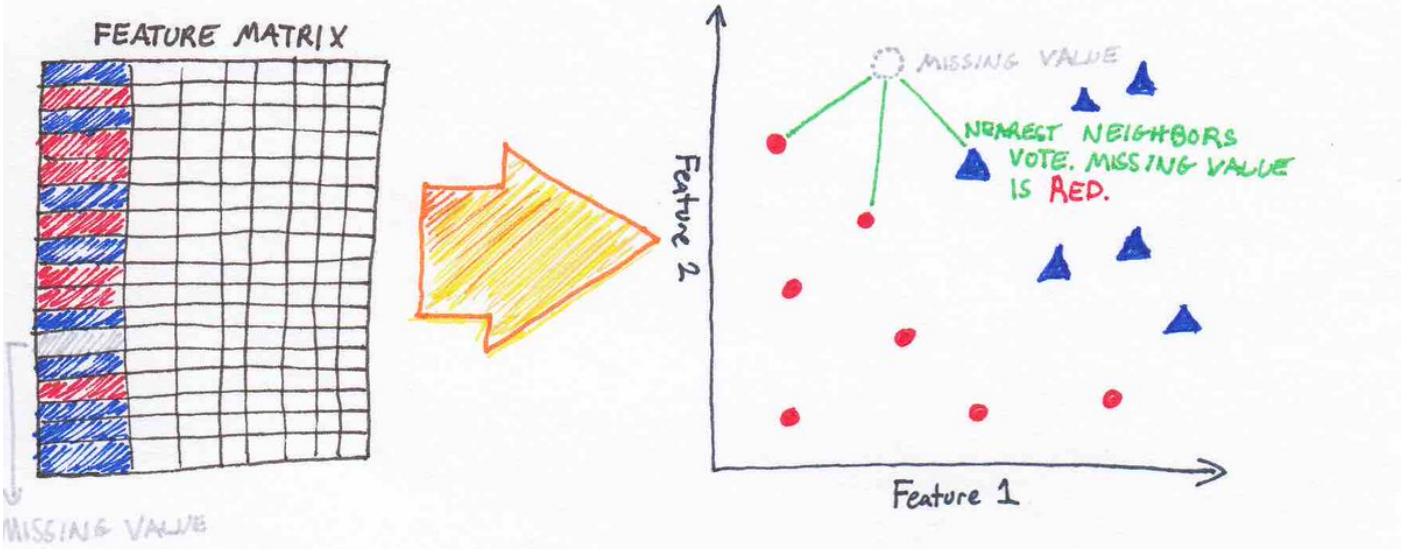
ChrisAlbon

IMPUTATION

USING K-NN



Imputation Using k-NN



K-NN has value. But has to calculate distance to every point which stops being scalabe in high N or/and high dimensionality.

BY CHRIS ALBON

IMPUTING

MISSING VALUES

1. If quantitative, replace with an average value.
2. If qualitative, replace with most common value.
3. Use a model to predict the missing values. For example, k-nearest neighbors.

ChrisAlbon

INITIALIZING WEIGHTS IN FEEDFORWARD NEURAL NETWORKS

- Initialize with small random numbers.
- Common to draw initial weights from normal distribution.
- Biases initialized as zero or small positive numbers.

Chris Albon

INTERACTION TERM

Interaction terms allow us model relationships when the effects of a feature on the target is influenced by another feature.

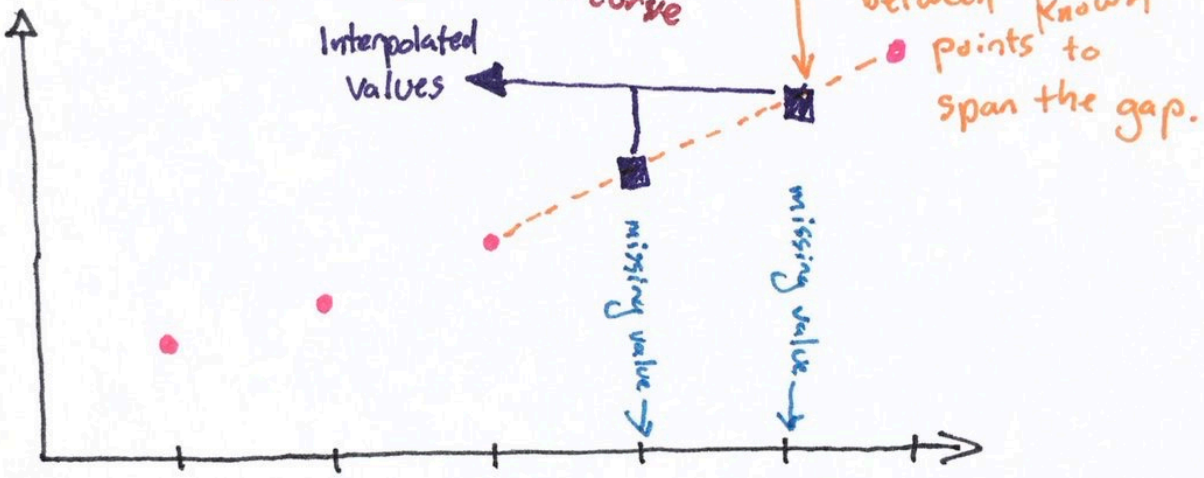
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

The interaction of features x_1 and x_2.

ChrisAlbon

Interpolation

Strategy to fill in gaps of missing values by drawing a line or curve between known values



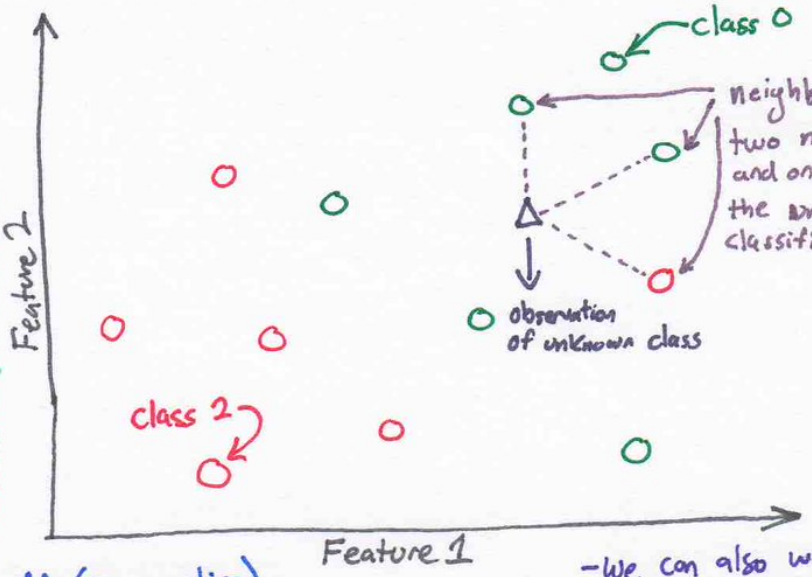
JACOBIAN MATRIX

When both the inputs and outputs of a function are vectors, the matrix containing all the first-order partial derivatives is called the Jacobian.

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Chris Albon

K-Nearest Neighbor



- Scaling is important
- The larger the neighborhood, the less variance and more bias
- KNN doesn't "learn" per-se. It is lazy and instead just memorizes the data.

- K should be odd (so no ties)
- Distance can be thought of as similarity
- Variety of distance metrics can be used.
- When we have categorical features 1 if same class and 0 otherwise. If we only have binary features we can use Hamming distance.

- We can also weight the voting by the distance to the neighbors. For example, closer observation's vote is worth more.

BY CHRIS ALBON

K-FOLD

CROSS-VALIDATION

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} Loss_i$$

- $CV_{(k)}$: # OF FOLDS
- $\frac{1}{k}$: # OF FOLDS
- $Loss_i$: FOR EXAMPLE:
  - MEAN SQUARED ERROR
  - LOG-LOSS
  - ACCURACY

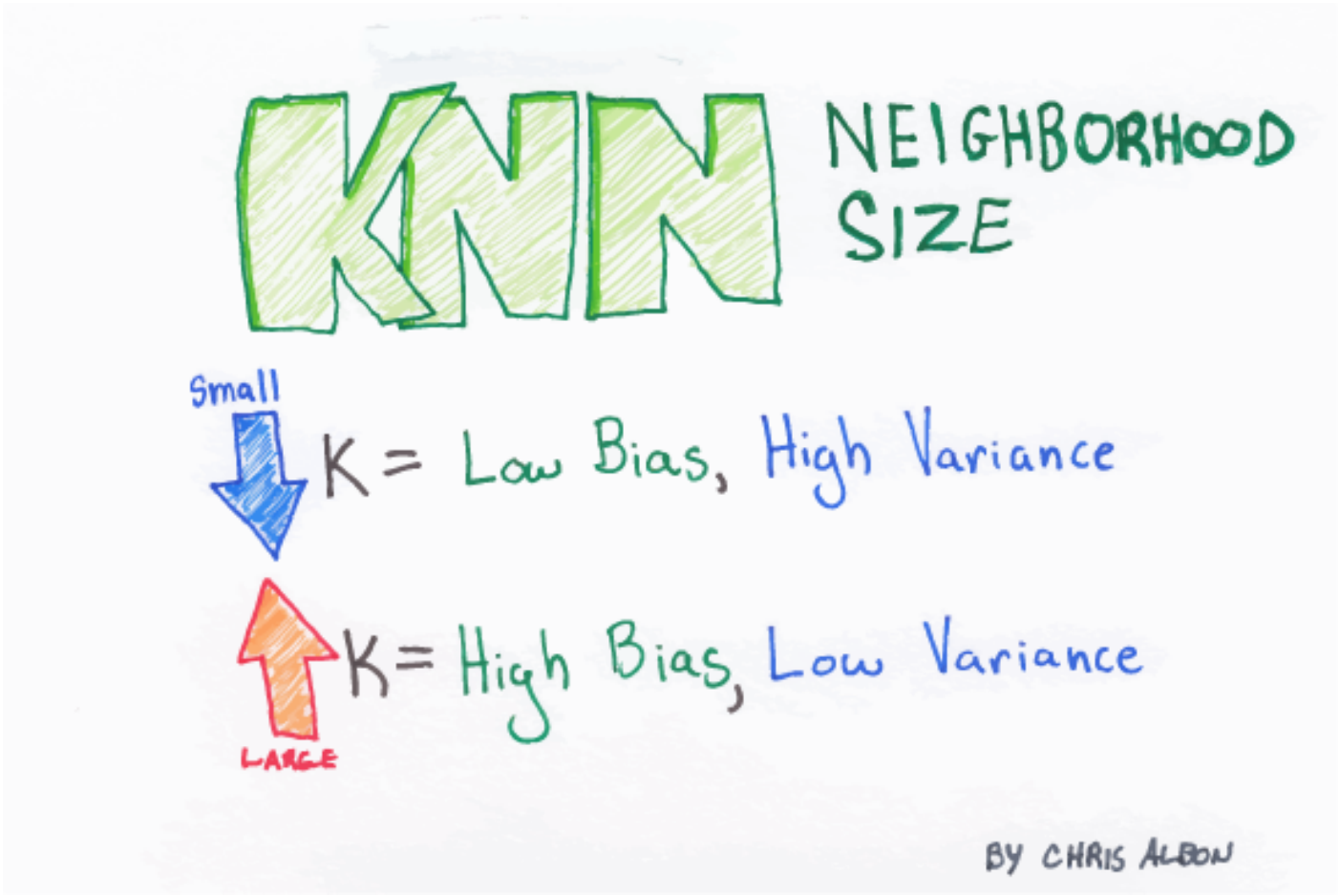
K=10 IS COMMON

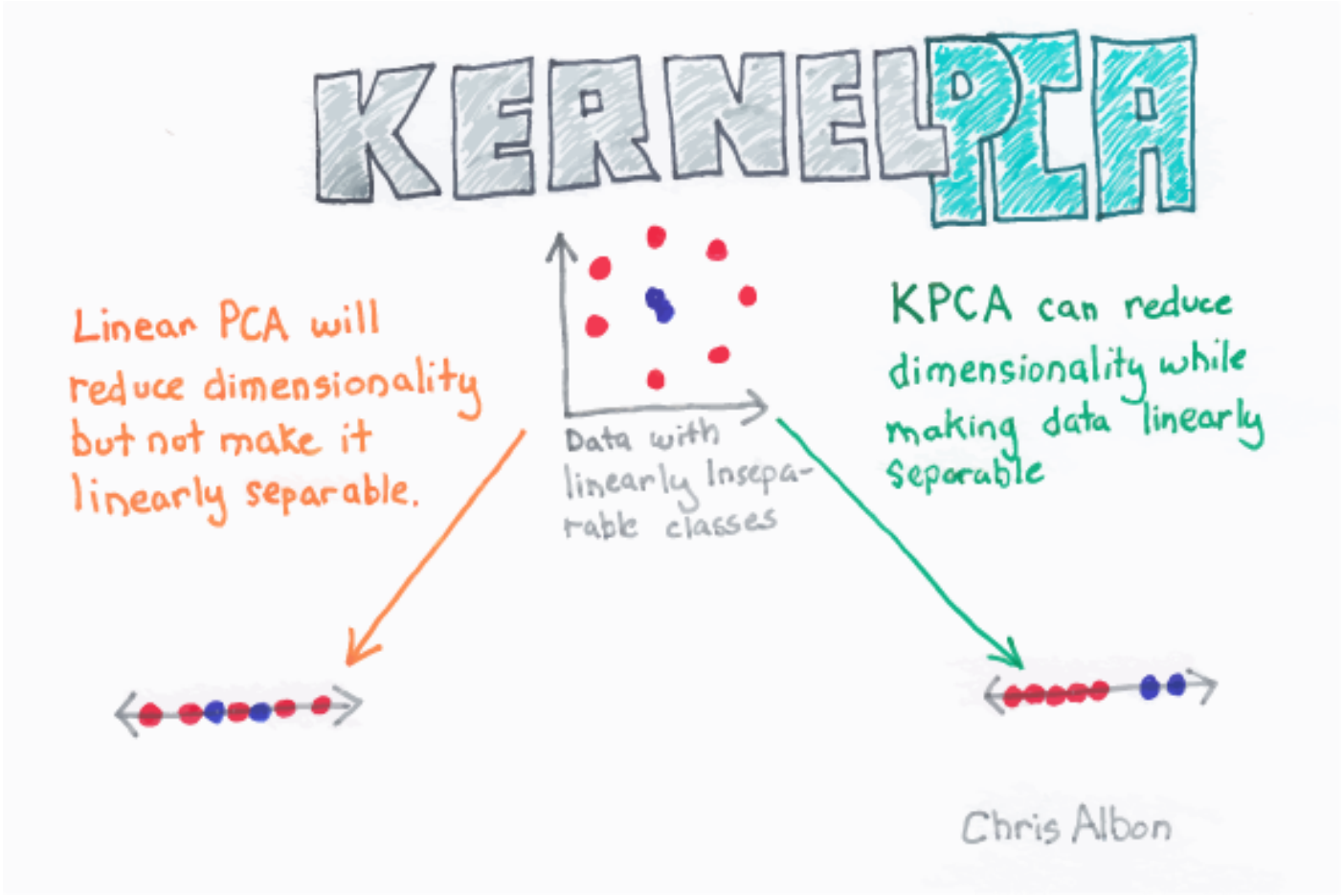
K-NEAREST NEIGHBORS

TIPS AND TRICKS

1. All features should use the same scale.
2. K should be odd to avoid ties.
3. Votes can be weighted by the distance to the neighbor so closer observations' votes are worth more.
4. Try a variety of distance measurements.

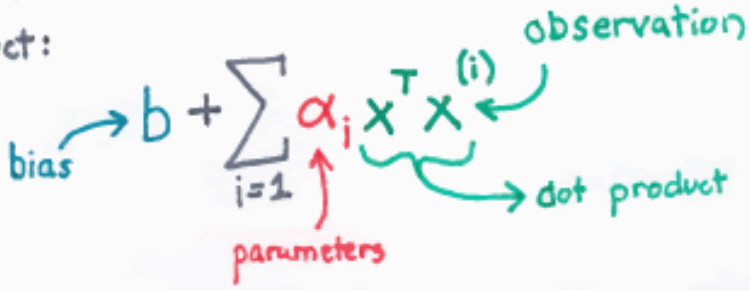
Chris Albon


KNN
NEIGHBORHOOD SIZE
Small
K = Low Bias, High Variance
K = High Bias, Low Variance
LARGE
BY CHRIS ALBON


KERNEL PCA
Linear PCA will reduce dimensionality but not make it linearly separable.
Data with linearly inseparable classes
KPCA can reduce dimensionality while making data linearly separable
Chris Albon

KERNEL TRICK

Support vector classifiers can be written as a dot product:



The kernel trick is to replace the dot product with a kernel:

$$b + \sum \alpha_i k(x, x^{(i)})$$ Kernel

Allows for non-linear decision boundaries and computational efficiency.

Chris Albon

K-nearest Neighbors

$$Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

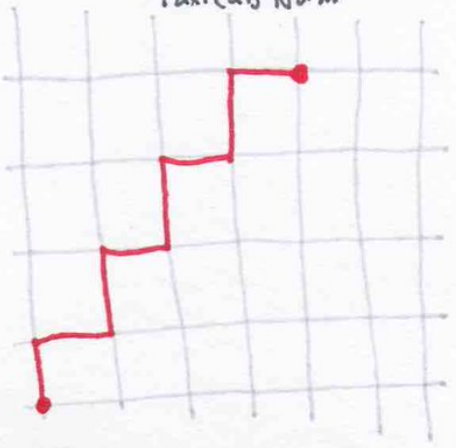
- $Pr(Y = j \mid X = x_0)$: $Y = j$ — category j; $X = x_0$ — features
- $\frac{1}{K}$: # of neighors
- $\sum_{i \in N_0}$
- $I(y_i = j)$: indicator; y_i — True y; j — category j

L1 Norm

(Manhattan Norm)

Also Called "Taxicab Norm"

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

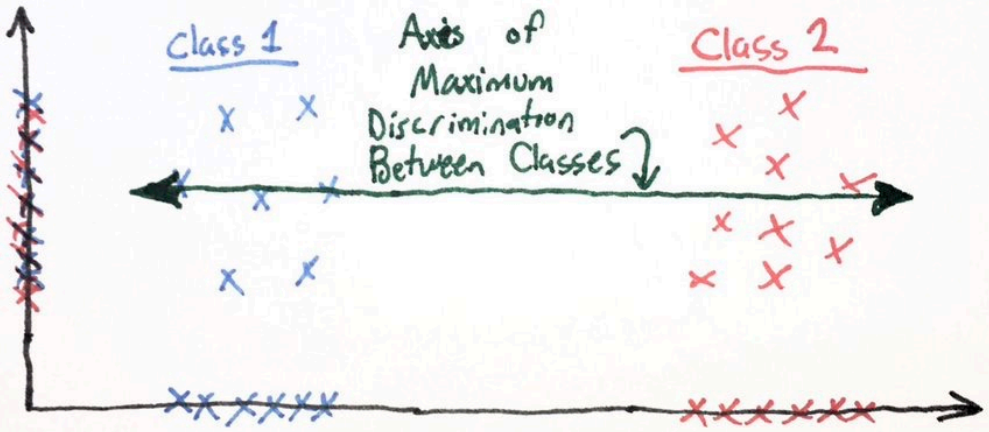


BY CHRIS ALBON

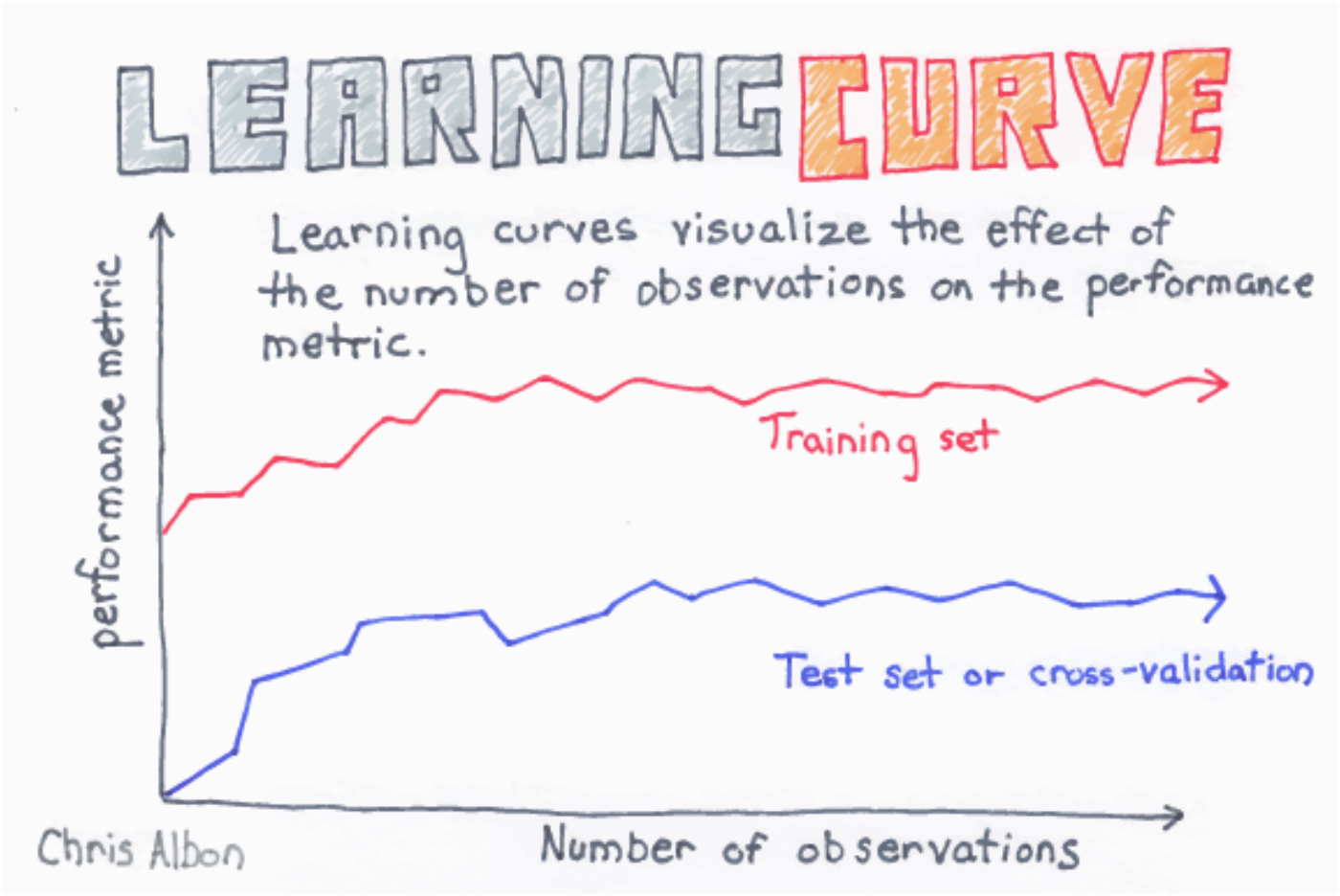
Why is nearest neighbor Lazy?

Because KNN doesn't learn so much as memorizes the data. That is, it doesn't learn any parameters.

Linear Discriminant Analysis For Dimensionality Reduction



Find Axes That Maximize Seperability Between Classes.
Project data down.


LEARNING CURVE
Learning curves visualize the effect of the number of observations on the performance metric.
performance metric
Training set
Test set or cross-validation
Chris Albon
Number of observations

Leave One Out Cross Validation

$$CV_{(n)} = \frac{1}{n} \sum_{i=}^{n} MSE_i$$

In LOOCV, you only have sample, so it is not really "Mean" squared error so much as just squared errors.

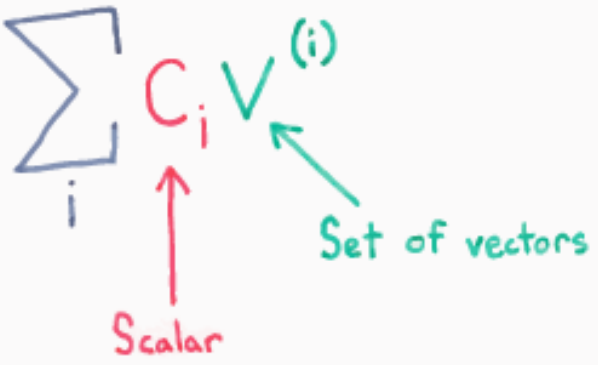
number of observation

Our score/loss metric can be anything but MSE is common in regression.

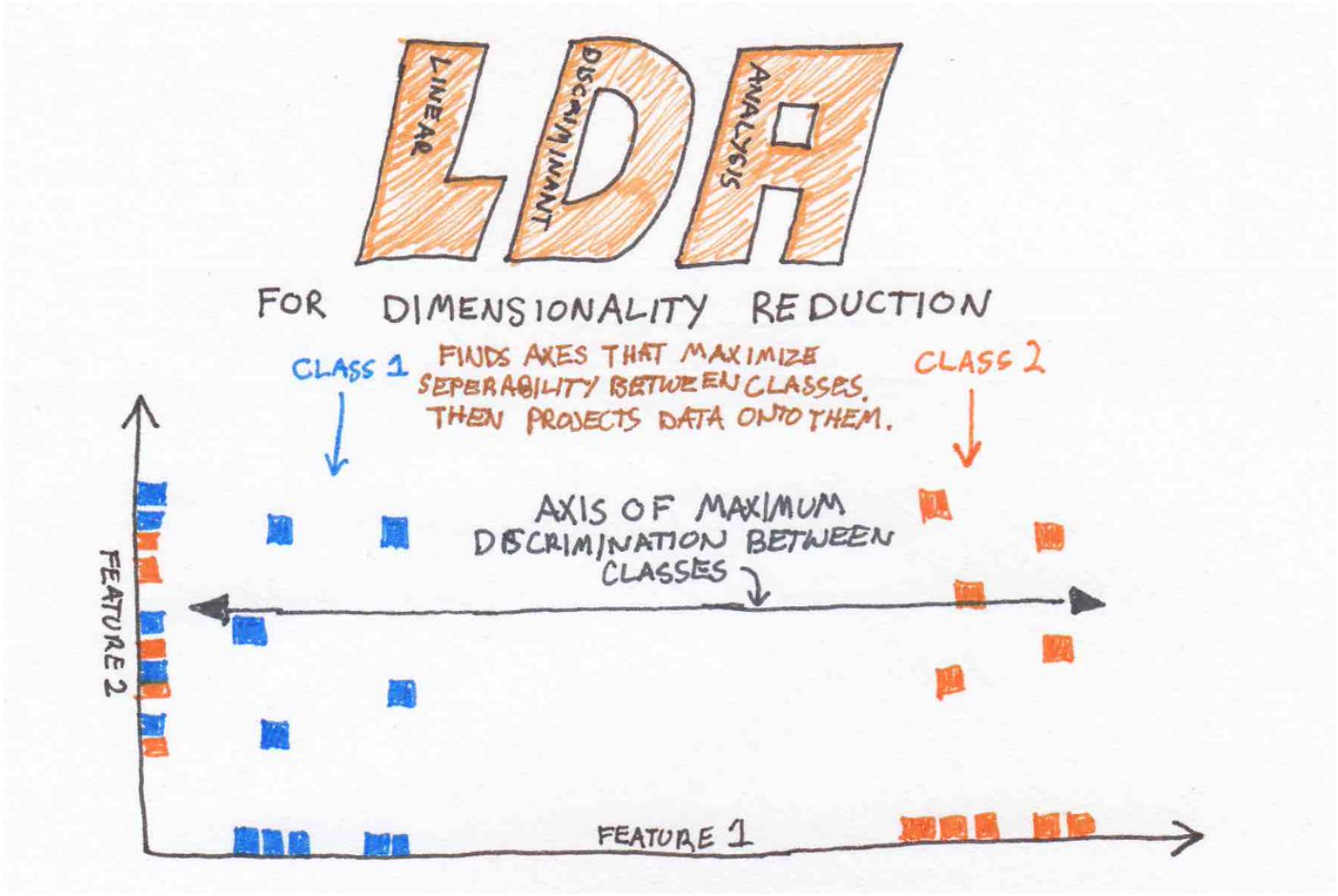
Thanks Sebastian Raschka (Buy his book)

LINEAR COMBINATION

OF A SET OF VECTORS



ChrisAlbon


LDA
LINEAR
DISCRIMINANT
ANALYSIS
FOR DIMENSIONALITY REDUCTION
CLASS 1
FINDS AXES THAT MAXIMIZE SEPERABILITY BETWEEN CLASSES, THEN PROJECTS DATA ONTO THEM.
CLASS 2
AXIS OF MAXIMUM DISCRIMINATION BETWEEN CLASSES
FEATURE 2
FEATURE 1

Linear Regression Problems and Solutions

1. Non-linearity of X-y relationship.
   - DIAGNOSIS: RESIDUAL PLOT
   - SOLUTION: NON-LINEAR TRANFORMATONS OF X
2. Correlation of error terms
   - DIAGNESIS: correlation
   - SOLUTION: EXPERIMENTAL DESIGN.
3. Non-constant error variance:
   - DIAGNOSIS: RESIDUAL PLOT
   - SOLUTION: CONCAVE TRANFORMATION LIKE LOG
4. Outliers:
   - DIAGNOSIS: RESIDUAL PLOT (STUDENTIZED) $\longrightarrow \frac{e_i}{\sigma_{e_i}}$ (e_i ← error, σ_{e_i} ← standard error of e_i)
   - SOLUTION: REMOVE. LOG.
5. High leverage points:
   - DIAGNOSIS: LEVERAGE STATSTIC $\longrightarrow h_i = \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum^{1} (x_i - \bar{x})^2}$
6. Collinearity:
   - DIAGNOSIS: VIF
   - SOLUTION: DROP VARIABLE

LOG-SUM-EXP

Imagine we want to calculate:

$$c = \log \sum_{i=1}^{n} e^{x_i}$$

If the scale of x is very small or large it can create a number, C, that is too small or large to represent on the computer. This is called underflow and overflow. The log-sum-exp trick exploits the fact that:

$$\log \sum_{i=1}^{n} e^{x_i} = a + \log \sum e^{x_i - a} \quad \text{where } a = \max(x)$$

This allows us to shift the center so the greatest value of $(x_i - a)$ is zero, preventing underflow and overflow.

MANHATTAN DISTANCE

$$\sum_{i=1}^{n} \left| p_i - q_i \right|$$

Example:

$Q = [3, 4, 5]$

$P = [0, 1, 2]$

MANHATTAN DISTANCE = $|0-3| + |1-4| + |2-5|$

MATRICES

Two dimensional array of scalars.

$$A_{m \times n} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$$

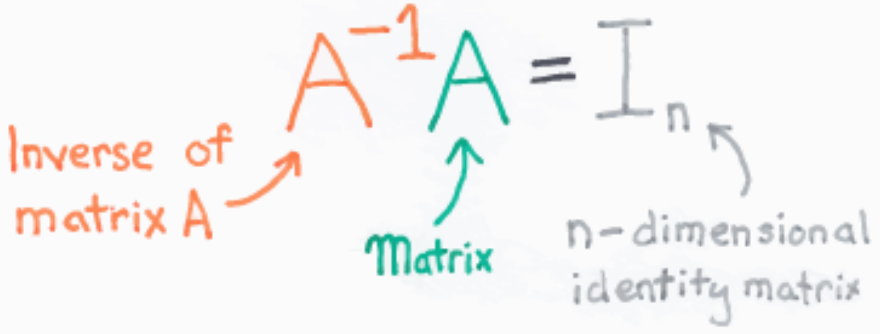
rows

columns

Scalar

ChrisAlbon

MATRIX INVERSE



ChrisAlbon

MAX NORM

The element of the vector with the largest absolute value.

$$\|X\| = \max_i |X_i|$$

$\|X\|$: vector X

$|X_i|$: value i

By Chris Albon

Mean Squared Error

Commonly used in regression.

$$MSE = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2$$

(y_i: true y; $\hat{f}(x_i)$: predicted y)

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + Bias\left(\hat{\theta}, \theta\right)^2$$

(n: number of observations)

MSE as the bias-variance trade-off

MEANSHIFT CLUSTERING

BY ANALOGY

Imagine a foggy football field with 100 people standing on it. Because of the fog, people can only see a short distance. Every minute each person looks around and takes a step in the direction of the most people they can see. As time goes on, people start to group up as they repeatedly take steps towards larger and large crowds The end result is clusters of people around the field.

ChrisAlbon

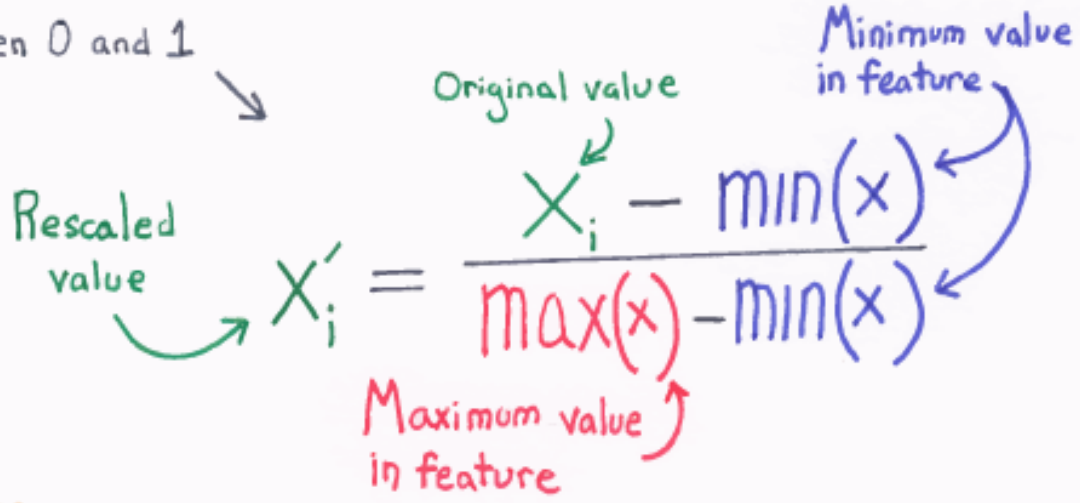
Min-Max Scaling

$$X = \frac{X - MIN}{MAX - MIN}$$

sklearn.preprocessing.MinMaxScaler.

MIN MAX SCALING

Rescales feature values to between 0 and 1



Chris Albon

MAR

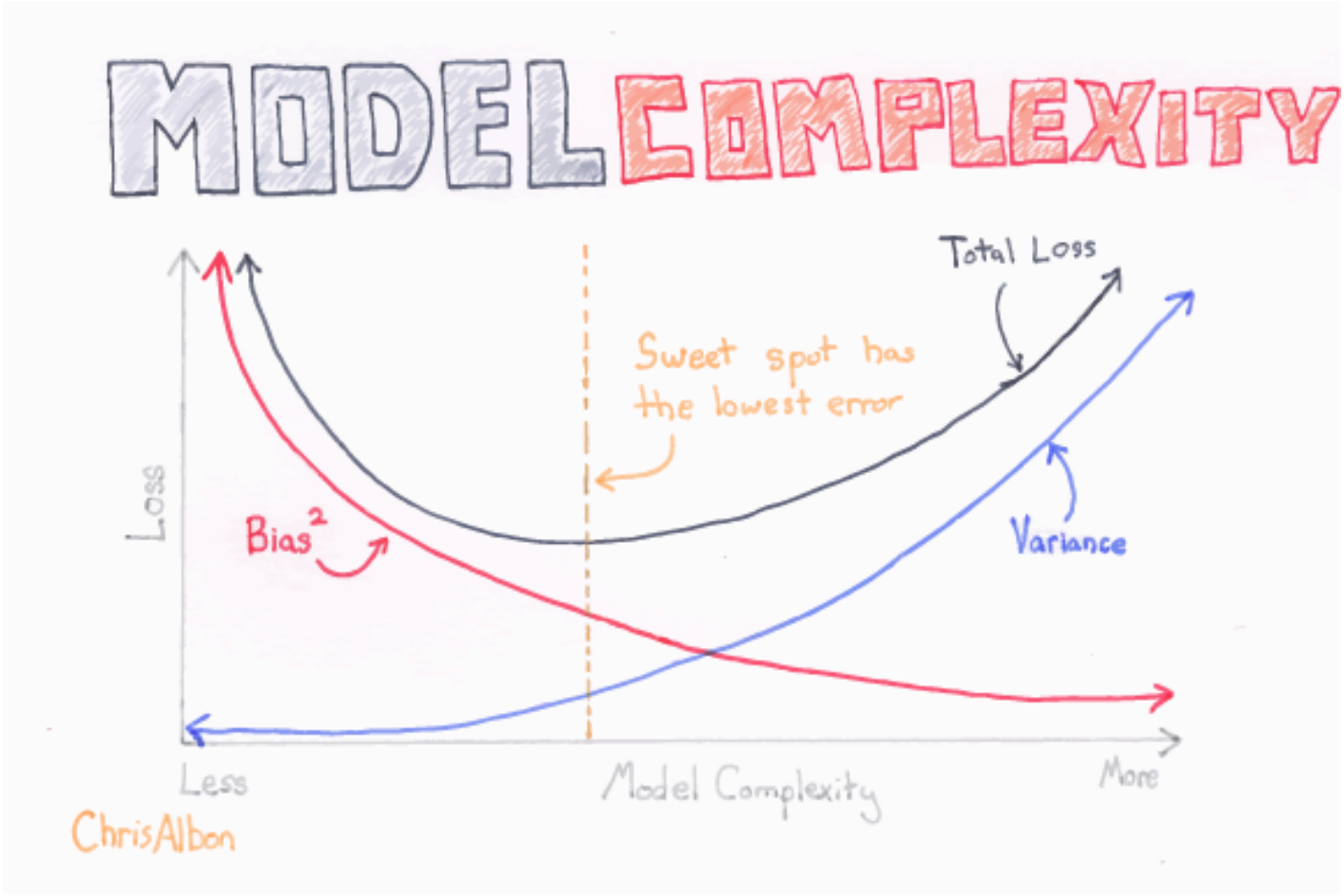
MISSING AT RANDOM

A type of missing data where the probability a value is missing is not completely at random but depends on information captured in other features.

For example, men are less likely to answer a salary question in a survey but we capture gender identity in another question.

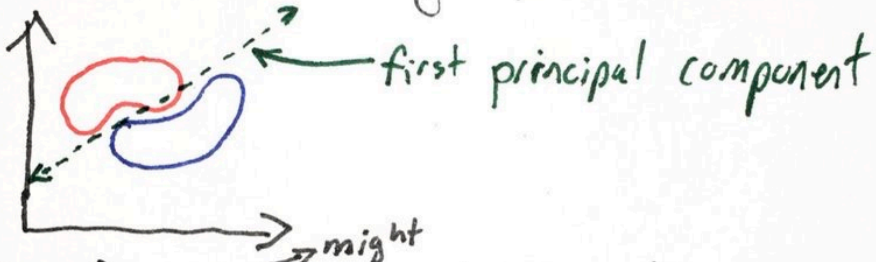
Chris Albon


MNAR
MISSING NOT AT RANDOM
A type of missing data where the probability a value is missing is not random and depends on information not captured in the other features.
For example, men are less likely to answer a salary question in a survey but we do not capture gender identity in another feature.
Chris Albon


MODEL COMPLEXITY
Total Loss
Sweet spot has the lowest error
Loss
Bias²
Variance
Less
Model Complexity
More
ChrisAlbon

Motivation For Kernal PCA

When data is not linearly separable:



PCA (linear) will might merge classes:

BAD!

But Kernal PCA can work:

GOOD!

MSE vs MAE

Benefits of Squaring:

- Always returns a positive value
- Emphasizes large errors ← important
- Continuosly differentiable
- Version of L2 norm

BY CHRIS ALBON

Multicolinearity

When collinearity exists between three or more coefficients

Test with VARIANCE INFLATION FACTOR (VIF)

$$\frac{\text{Var}(\hat{B}_j) \text{ when part of the whole model}}{\text{Var}(\hat{B}_j) \text{ when only variable in model}}$$

VIF of 5 or more is considered bad.

NAIVE BAYES CLASSIFIER!

$$p(\text{class} \mid \text{data}) = \frac{p(\text{data} \mid \text{class}) \times p(\text{class})}{p(\text{data})}$$

Normal distribution →

$$\frac{1}{\sqrt{2\pi\,\text{Var}(\text{male height})}} e^{-\frac{(x_i - \bar{x})^2}{2\pi\,\text{Var}(\text{height})}}$$

$$\text{posterior}(\text{male}) = \frac{p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male}) \times p(\text{male})}{\text{marginal probability}}$$

We ignore this, only compare numerators of the posterior.

BY CHRIS ALBON

NATURAL LOG

$$\ln\left(e^{x}\right) = x$$

Euler's number: 2.71828...

BY CHRIS ALBON

NON-PARAMETRIC METHODS

Non-parametric methods do not assume a functional form of the relationship between X and y.

ADVANTAGE: Can fit a wider ranges of types of relationships.

DISADVANTAGE: Often a large number of observations required to fit a model of any quality.

BY CHRIS ALBON

Normalization

Rescales the values of individual observation so that they have unit (1) norm.

Note: A lot of ambiguity around "normalization". Use scikit's lexicon.

Option 1: Euclidean Norm

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 \ldots x_n^2}$$

Option 2: Manhatten Norm

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Seen in text classification

NOTATION 1

$\neg$ Not

$\ni$ Such that

$a \in b$ a is an element of b

$\therefore$ Therefore

Chris Albon

NOTIONS OF PROBABILITY

Frequentist: The rate at which an event will occur if we repeat the experiment.

Bayesian: The degree of belief that something is true.

Chris Albon

Odds and Log Odds in Logistic Regression

ODDS: $\frac{p(x)}{1-p(x)} = e^{B_0 + B_1 X_1 + \ldots + B_n X_n}$

LOG ODDS: $\log\left(\frac{p(x)}{1-p(x)}\right) = B_0 + B_1 X_1 + \ldots B_n$

ODDS

$$\frac{Pr(y)}{Pr(\sim y)}$$

event

non-event

Odds is the ratio of the probability an event occurs with the probability of an event not occuring.

ChrisAlbon

One-Hot Encoding

Feature
Apple
Pear
Apple
Pear
Apple

Apple	Pear
1	0
0	1
1	0
0	1
1	0

nominal

One-Hot encoding allows us to turn categorical data into features with numerical values (that many ML algorithms need) while not mathematically imply any ordinal relationship between the classes.

Note:

Spliting one feature into many binary features can artificially reduce the feature importance in random forest because info is spread over many features.

OUT-OF-BAG ERROR

Method of estimating test error in bagged models. OOB error is the mean error of all observations, x_i, using only the trees which did not use x_i in their bootstrapped datasets. OOB error is as accurate as a test set of equal size as the training set (Breiman 1996).

BY CHRIS ALBON

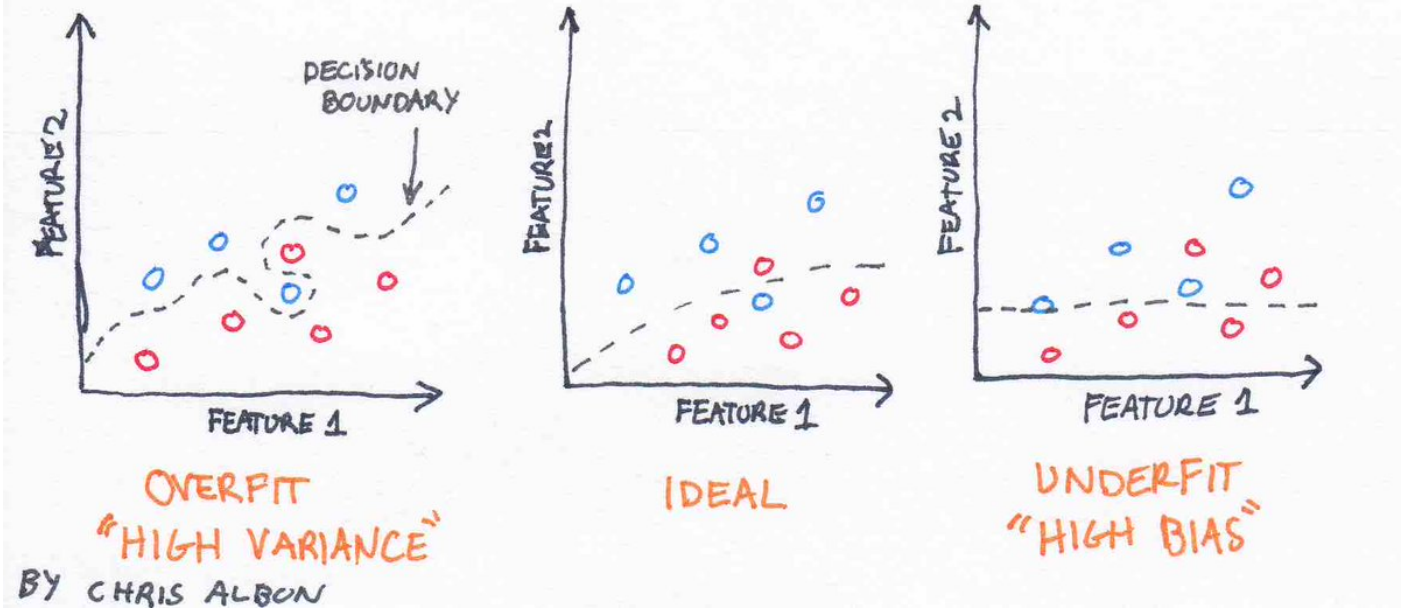
OUTLIARS

DROP
- Not a great option. We lose lots of info.
- Find out if genuine extreme value or broken sensor.

MARK
- Safest option. We can see if the
- outliars had an effect.

RESCALE
- Log values so outliars
- don't have as great an effect.

OVERFIT VS UNDERFIT



BY CHRIS ALBON

Parameters vs. Hyperparameters

PARAMETERS: LEARNED FROM THE TRAINING. FOR EXAMPLE: REGRESSION COEFFICIENTS.

HYPERPARAMETERS: SET BEFORE THE TRAINING STEP. "TUNED" THROUGH HYPER PARAMETER GRID SEARCH AND RELATED METHODS.

Parametric Methods

(1) Assume a functional form of f. For example: we can assume that f is linear.

(2) Use a method (OLS, MLE, etc) to train the model.

ADVANTAGE: If we choose the correct functional form, we reduce the estimation problem to estimating a small set of parameters.

Pearson's Correlation Coefficient

$$\rho = \frac{Cov(X, Y)}{S_x S_y}$$

← covariance

← standard deviations

GOOD IF: Variables are roughly normal or if the relation of the variables is roughly linear.

BAD IF: Outliars present or skewed distribution.

CORRELATION

(Pearson's R)

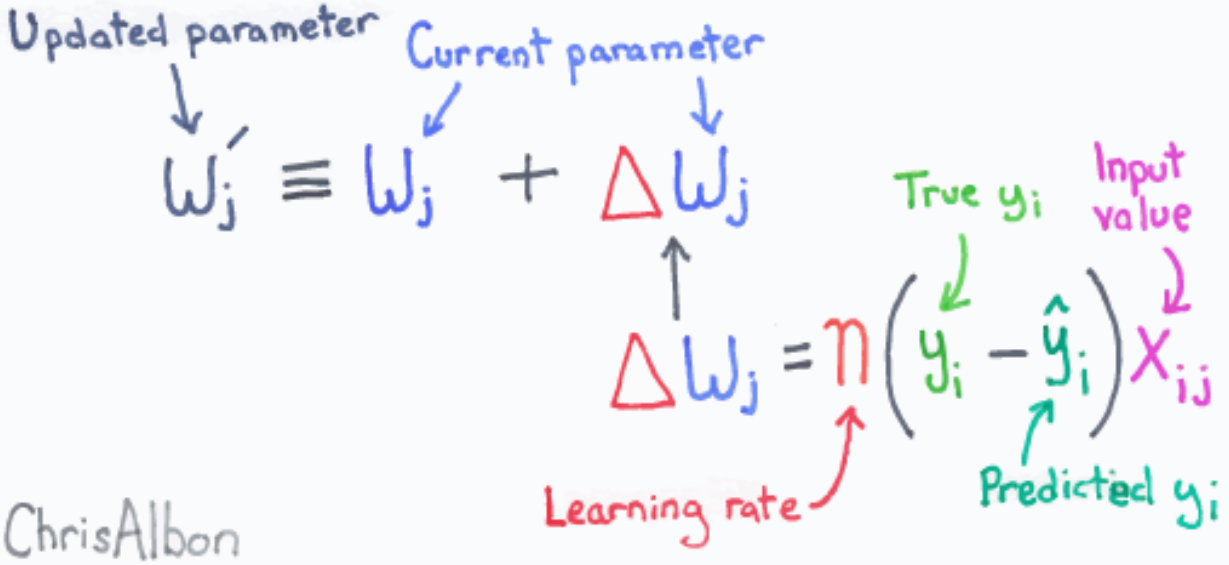
value of x_i — mean of x — value of y_i — mean of y

$$Cor(X,Y) = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}\sqrt{\sum \left(y_i - \bar{y}\right)^2}}$$

Ranges between -1.0 and +1.0. The closer to 0.0 the less linear dependence between variables.

BY CHRIS ALBON

PERCEPTRON LEARNING



ChrisAlbon

Polynomial Regression

$$\hat{y}_i = B_0 + B_1 x_i + B_2 x_i^2 \ldots e_i$$

ADDS NON-LINEARITY
WHILE STILL OBEYING THE
RULES OF A LINEAR FUNCTION.

POWER RULE

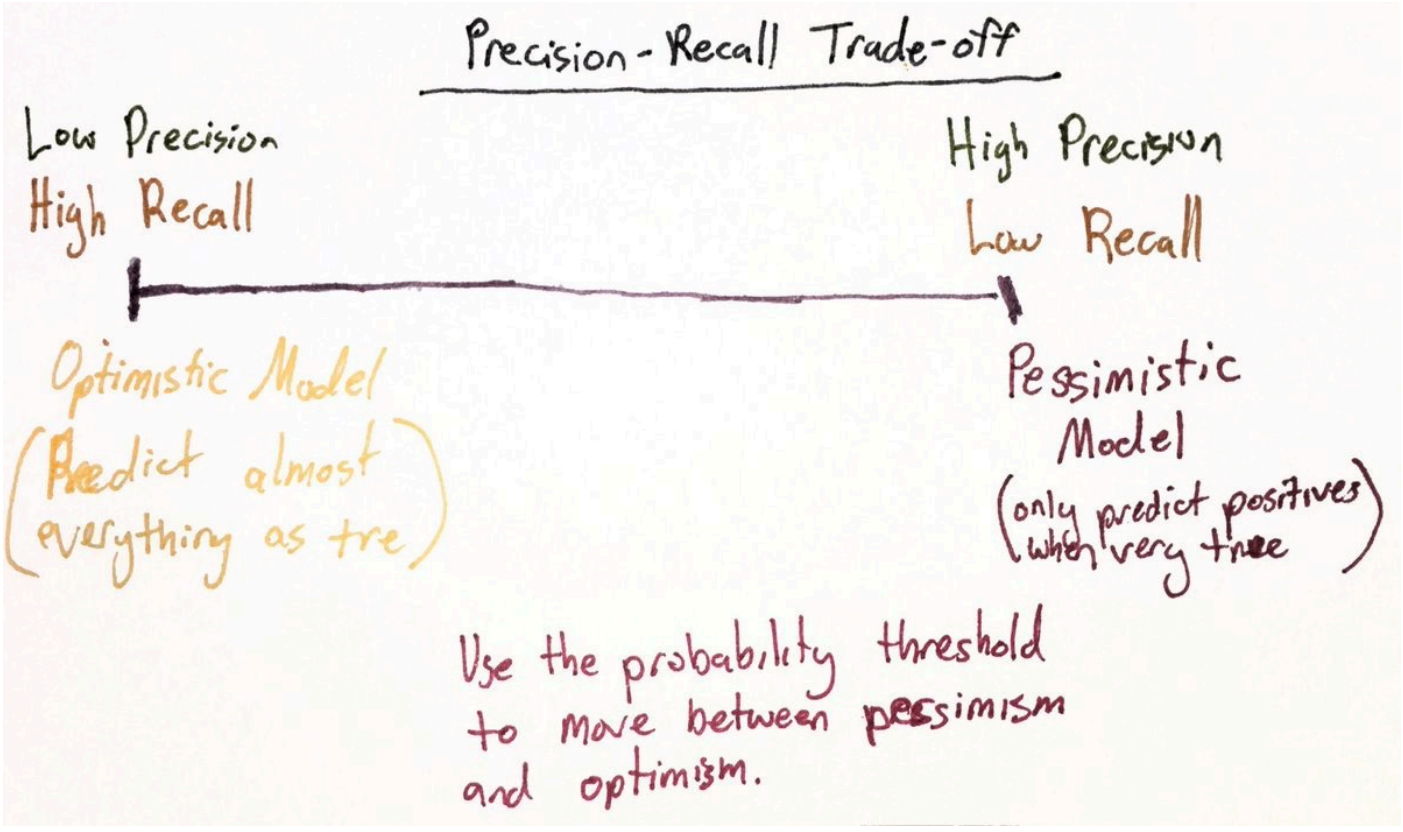
What is $\frac{d}{dx}x^3$? Use the power rule:

$$\frac{d}{dx}x^n = nx^{n-1}$$

Therefore the derivative is $3x^2$

Chris Alban

Precision - Recall Trade-off



Use the probability threshold to move between pessimism and optimism.

Precision

"Precision is about the predicted positives"

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{Everything Predicted Positive}}$$

Precision is the ability of a classifier not to label as positive an observation that is negative. It measures the purity of positive predictions.

BY CHRIS ALBON

PREPROCESSING TEST + TRAINING SETS

1. Fit preprocessor to the traing set ONLY.
2. Apply it to both the training and test set.

WHY? Because we have to "pretend" the test set is unknown data.

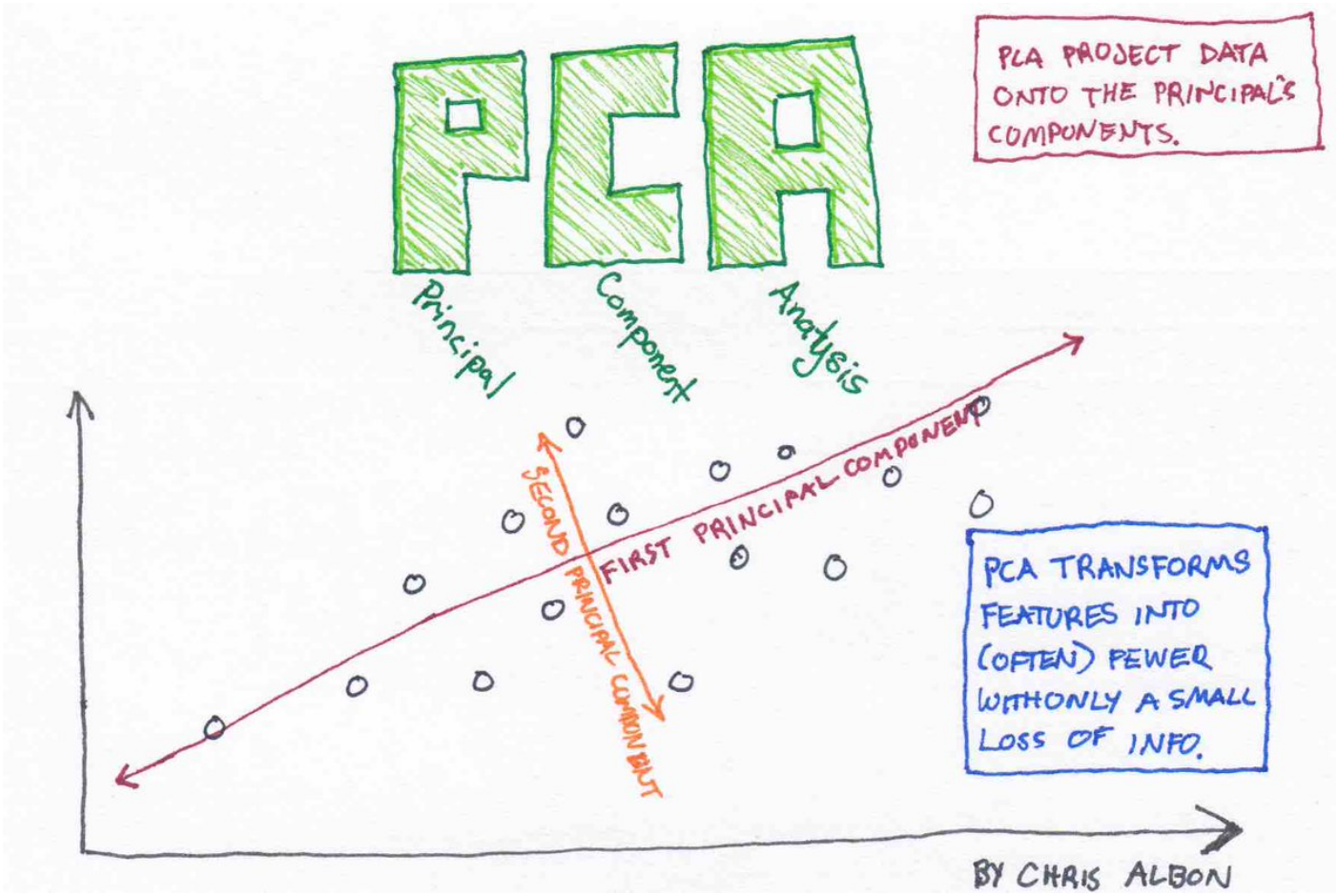
BY CHRIS ALBON

Preprocessing When Tests and Training.

1) Fit the preprocessor to the training set.

2) Apply it to both the training AND test set.

Why? Because we have to pretend the test set is new data.


PCA
Principal
Component
Analysis
PCA PROJECT DATA ONTO THE PRINCIPAL'S COMPONENTS.
FIRST PRINCIPAL COMPONENT
SECOND PRINCIPAL COMPONENT
PCA TRANSFORMS FEATURES INTO (OFTEN) FEWER WITHONLY A SMALL LOSS OF INFO.
BY CHRIS ALBON

PRINCIPAL COMPONENTS

Principal components are the linear combination of features that have the maximum variance out of all linear combinations.

Alternative interpretation: Principal components are low dimensional linear surfaces closest to the observations.

ChrisAlbon

PDF

Probability Density Function

The PDF is the probability distribution of a continuous random variable. PDFs tell us the probability of an infinitely small region. We can use integration to find the probability

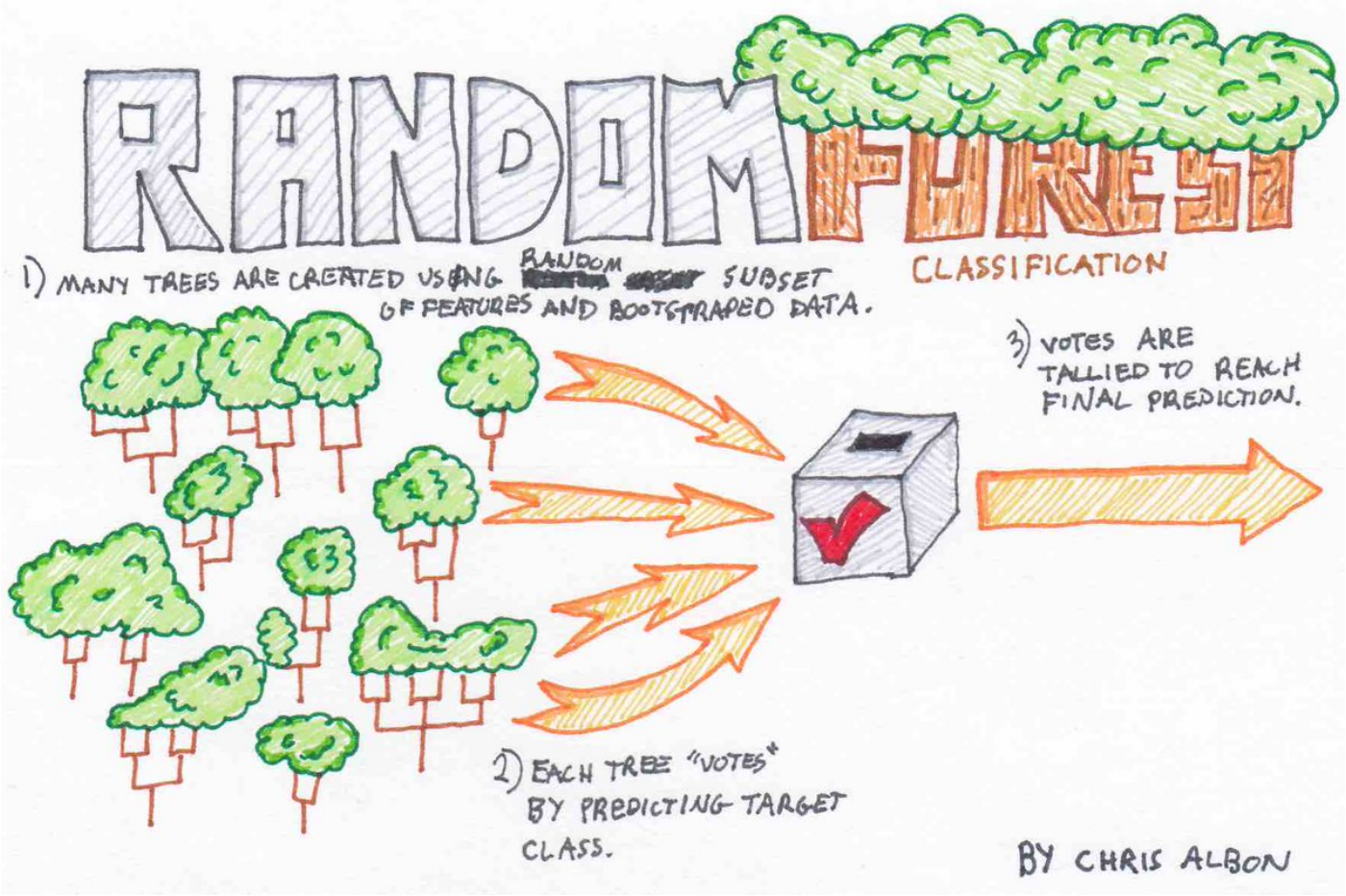
R^2

R^2 looks at How much variance in the target vector is explained by the features

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- y_i: TRUE y
- $\hat{y}_i$: PREDICTED y
- $\sum (y_i - \hat{y}_i)^2$: VARIANCE IN PREDICTIONS VS. TRUE ys
- $\sum (y_i - \bar{y})^2$: VARIANCE IN TARGET VECTOR
- y_i: TRUE y
- $\bar{y}$: MEAN TRUE y

BY CHRIS ALBON

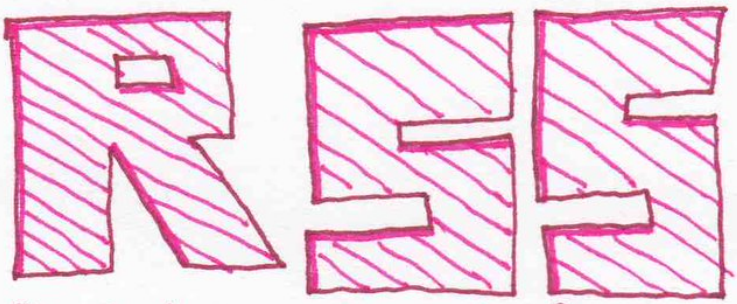

RANDOM FORESTS
CLASSIFICATION
1) MANY TREES ARE CREATED USING RANDOM SUBSET OF FEATURES AND BOOTSTRAPED DATA.
2) EACH TREE "VOTES" BY PREDICTING TARGET CLASS.
3) VOTES ARE TALLIED TO REACH FINAL PREDICTION.
BY CHRIS ALBON

"Recall is about the real positives"

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is the ability of the classifier to find all positive Examples. If we wanted to be certain to find all positive examples, we'd maximize recall.

BY CHRIS ALBON



Residual Sum of Squares

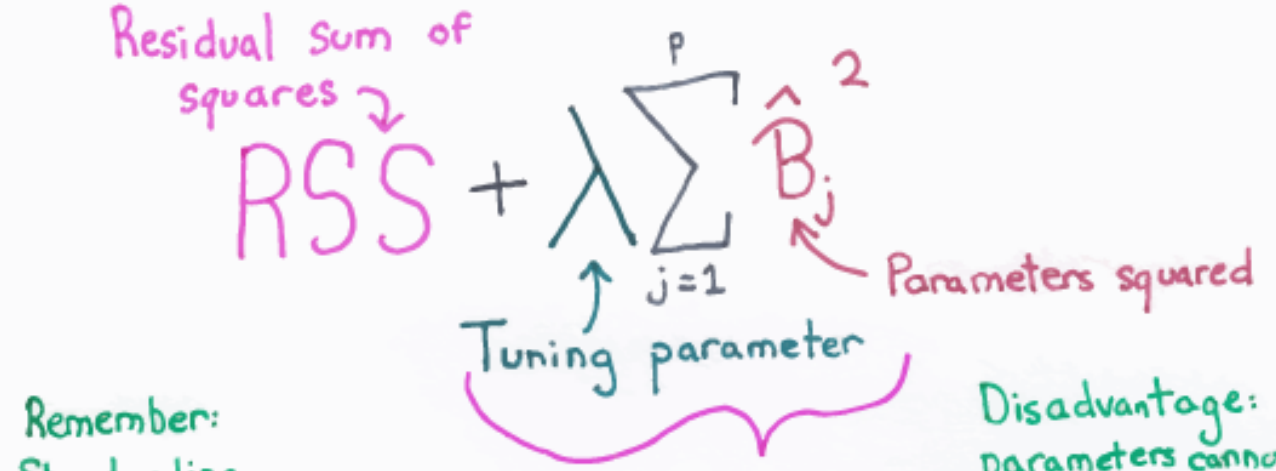
$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

y_i: true y_i

$\hat{y}_i$: predicted y_i (Also called "y-hat")

Note that squaring the error with more heavily penalizes a few large errors over many small errors even if the sum of errors are the same

RIDGE REGRESSION



Remember: Standardize the data first.

Chris Albon

Disadvantage: parameters cannot be zero like with Lasso regression.

RSS

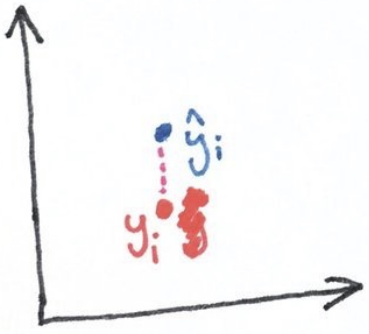
Residual Sum Of Squares

$$RSS = \sum_{i=1}^{n} \left(y_i - X_i^T B \right)^2$$

← Removes negative signs, penalizes larger errors more.

y_i = true y value

$\hat{y}_i$ = predicted y

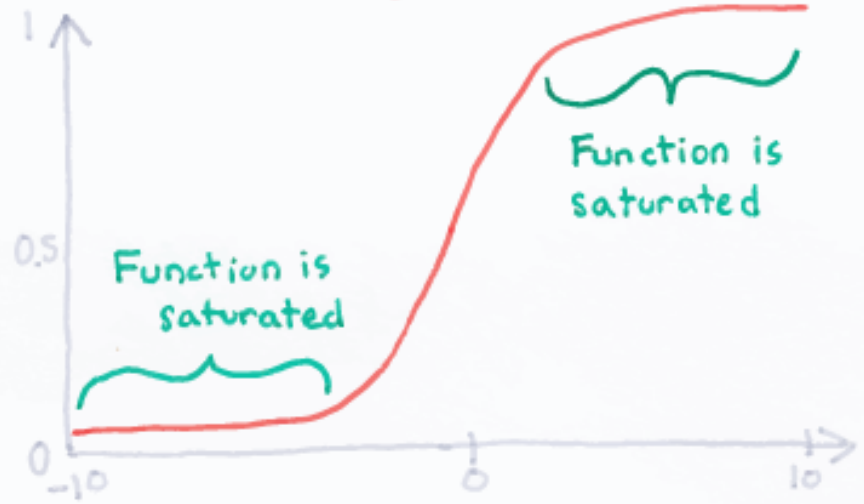


example!

y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	RSS
2	2	$0^2 = 0$	
4	2	$2^2 = 4$	8
-4	-2	$-2^2 = 4$	
3	1	$2^2 = 4$	

SATURATION

When a function's output is very insensative to inputs. Example: sigmoid function.



Chris Albon

Selecting # of components in PCA

Select smallest number of components such that:

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \| x_i - x_i \text{ approx} \|^2}{\frac{1}{m}\sum_{i=1}^{m} \| x_i \|^2} = 0.01$$

- $\frac{1}{m}\sum_{i=1}^{m} \| x_i - x_i \text{ approx} \|^2$ ← AVERAGE SQUARE PROJECTION ERROR
- DISTANCE BETWEEN X AND ITS PROJECTIONS.
- $\frac{1}{m}\sum_{i=1}^{m} \| x_i \|^2$ ← TOTAL VARIATION IN DATA
- $= 0.01$ ← 99% of VARIANCE RETAINED

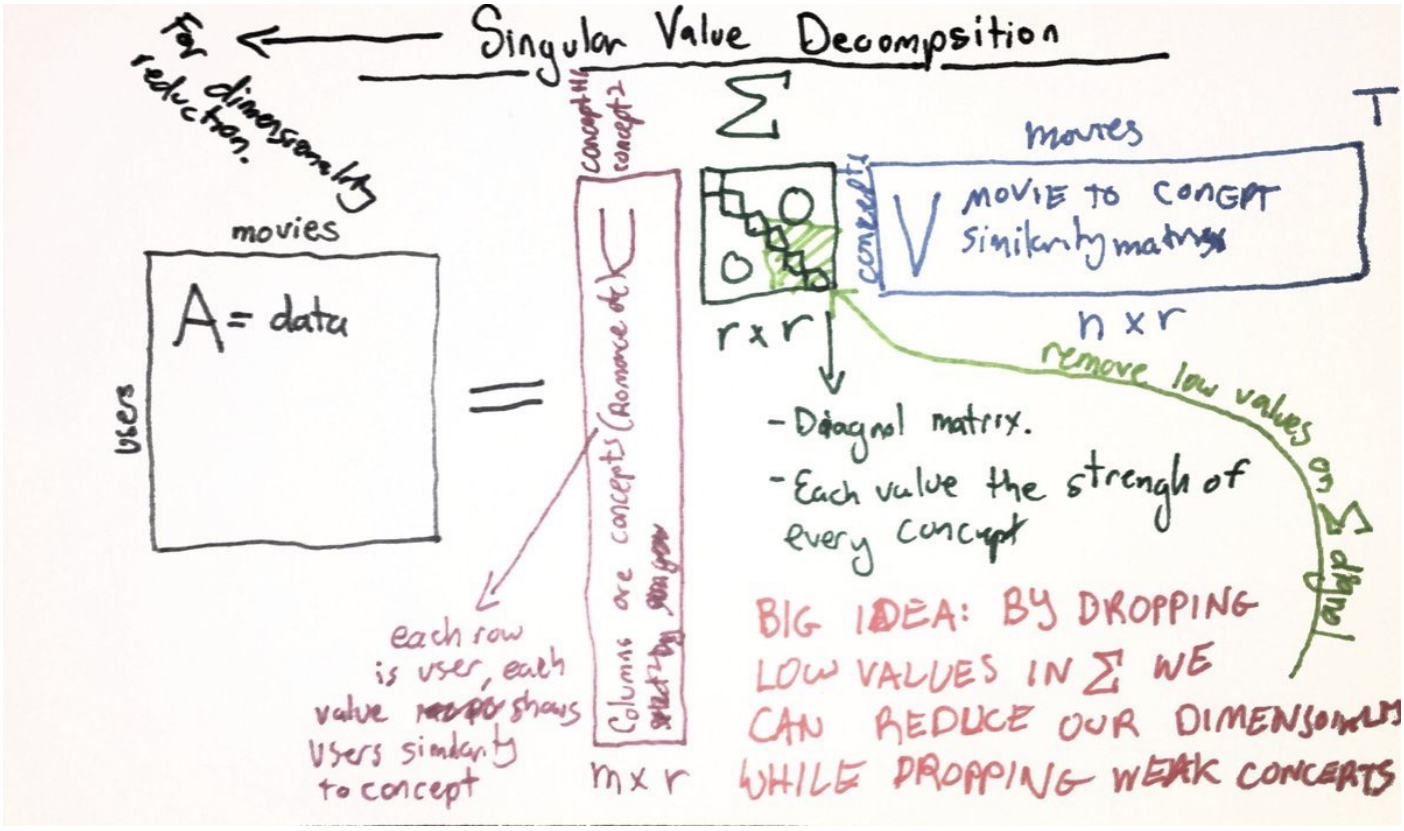
Sensitivity

Recall on real positive examples.

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Everything predicted positive correctly

Everything actually positive

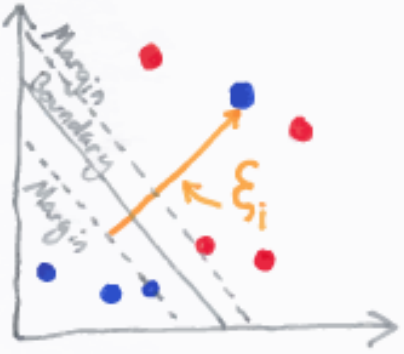

Singular Value Decomposition
for dimensionality reduction.
movies
A = data
users
=
concept #1
concept-2
Columns are concepts (Romance etc)
each row is user, each value shows users similarity to concept
m x r
Σ
r x r
- Diagonal matrix.
- Each value the strength of every concept
concepts
V
movies
MOVIE TO CONCEPT similarity matrix
n x r
T
remove low values or 0
BIG IDEA: BY DROPPING LOW VALUES IN Σ WE CAN REDUCE OUR DIMENSIONALITY WHILE DROPPING WEAK CONCEPTS

SLACK VARIABLE IN SOFT-MARGIN SVM

$\xi_i = 0$, the observation is correctly classified

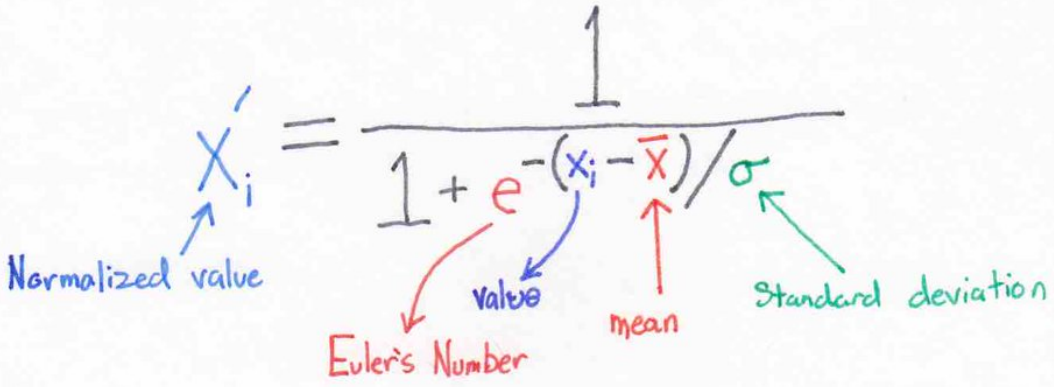
$0 < \xi_i < 1$, the observation is inside the margins

$\xi \geq 1$, the observation is misclassified



SOFTMAX NORMALIZATION

Reduces the influence of outliers without having to drop them.



BY CHRIS ALBON

SPARSITY

Sparse matricies allow for more efficient data storage, particularly when most values are zero or missing.

$$\text{Sparsity} = \frac{\#\text{ of zero/missing values}}{\text{Total \# of values}}$$

$$\text{Sparsity} = 1 - \text{density}$$

BY CHRIS ALBON

SQUARE ROOT

Principle Root

$$\sqrt[r]{x} = x^{\frac{1}{r}}$$

Square Root

$$\sqrt[2]{x} = x^{\frac{1}{2}}$$

Often used to reverse or undo a squaring. For example:

$$\sqrt{x^2} = x$$

BY CHRIS ALBON

Standard Error

ANSWERS: TELLS US THE AVERAGE AMOUNT $\hat{\mu}$ DIFFERS FROM THE TRUE μ.

$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

$\hat{\mu}$: sample mean

$\hat{\mu}$: sample mean

σ^2: variance $= \frac{\sum (x-\mu)^2}{N}$

n: # of observations

unit is the standard deviation

Standardization
(z-score scaling)

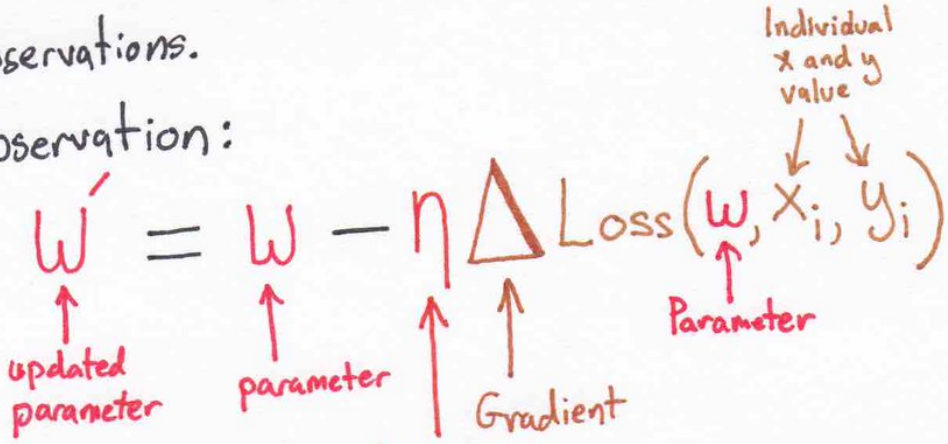
Rescales input data (features) to be approximently standard norally distributed.

$$x' = \frac{x - \bar{x}}{\sigma}$$

- x': Scaled value
- x: value
- $\bar{x}$: mean value of feature
- σ: standard deviation

STOCHASTIC GRADIENT DESCENT

1. Shuffle Observations.
2. For each observation:



3. Repeat until ~minimum value achieved.

BY CHRIS ALBON

STOP WORDS

Any word to remove before processing. Frequently Stop words are extremely common words with little informational value.

Examples
- it
- me
- myself
- we
- the
- and

Chris Albon


STRATEGIES WHEN YOU HAVE
HIGH VARIANCE
- Weight Decay
- Drop out or bagging
- Dimensionality reduction
- Feature selection
Chris Albon

SUPERVISED VS. UNSUPERVISED

In supervised learning, for every observation we have some feature values X_i and some target vector, y_i. We use both to train a model that take in some x_i values and outputs a predicted y_i.

In unsupervised learning, we only have features, and do not have a target vector. This makes the estimation problem much more difficult. When possible, use supervised learning.

BY CHRIS ALBON

SUPERVISED VS UNSUPERVISED

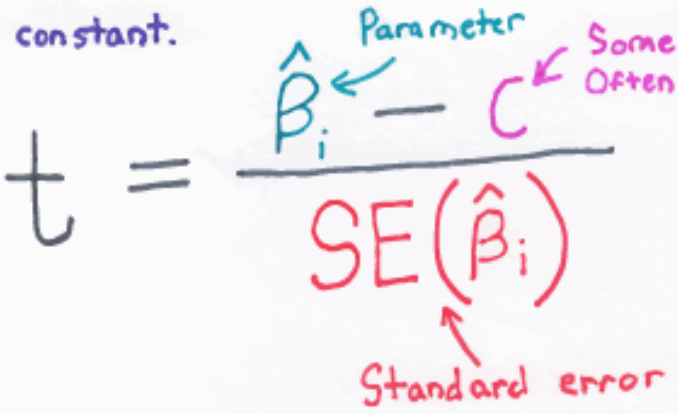
In supervised learning, for every observation we have some feature values and some target vector or tensor. We use both to train a model that takes in some x values and outputs a predicted value.

In unsupervised learning, we only have features and do not have a target. This makes the estimation problem more difficult. When possible, use supervised learning.

Chris Albon

T-STATISTIC

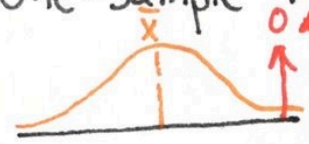
Number of standard deviations a parameter is away from a constant.



Chris Albon

T - Tests

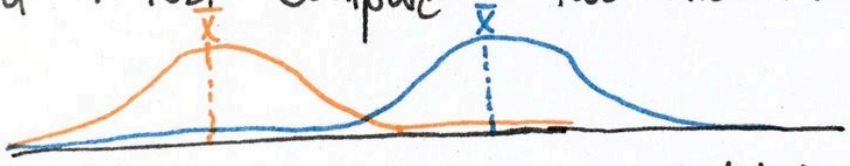
One-sample t-test: Compare sample to some constant.



$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

t ← t-statistics, $\bar{x}$ ← mean, μ_0 ← some value, s ← standard deviation, n ← # of obs

Unpaired t-test: Compare if two distributions are different



Paired t-test: Compare if two distributions of repeated sampling are the same.

TF-IDF

THIS IS THE NOTATION USED BY SCIKIT-LEARN

number of times term appears in a doc.

# of documents where a word appears

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

term frequency

# of times word appear in a doc

inverse document frequency

number of documents

$$\log \frac{1 + n_d}{1 + df(d, t)} + 1$$

document frequency of the term, t.

BY CHRIS ALBON

The Argument For Parametric Models

- When the data generating function roughly matches a parametric probability distribution we can limit our calculations to only its parameters (mean, variance, etc.). This lets us know (assume) a lot using only a little information.
- The flexibility of many probability distributions means that often the choice of the functional family of distribution (normal vs. student etc.) is often not a problem
- Range matching is, however, important. If we want a probability, we should not choose a probability distribution that outputs numbers greater than 1, for example.

BY CHRIS ALBON

THE EFFECT OF DROPOUT ON HIDDEN UNITS

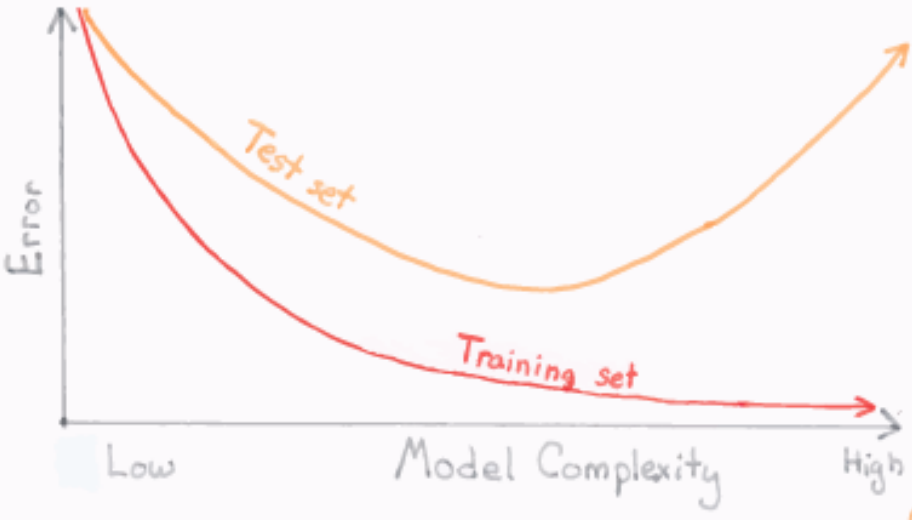
In dropout, each hidden unit must learn to perform well regardless of the other units in the network. The learned robustness helps the network perform well in the face of unseen test data.

Chris Albon

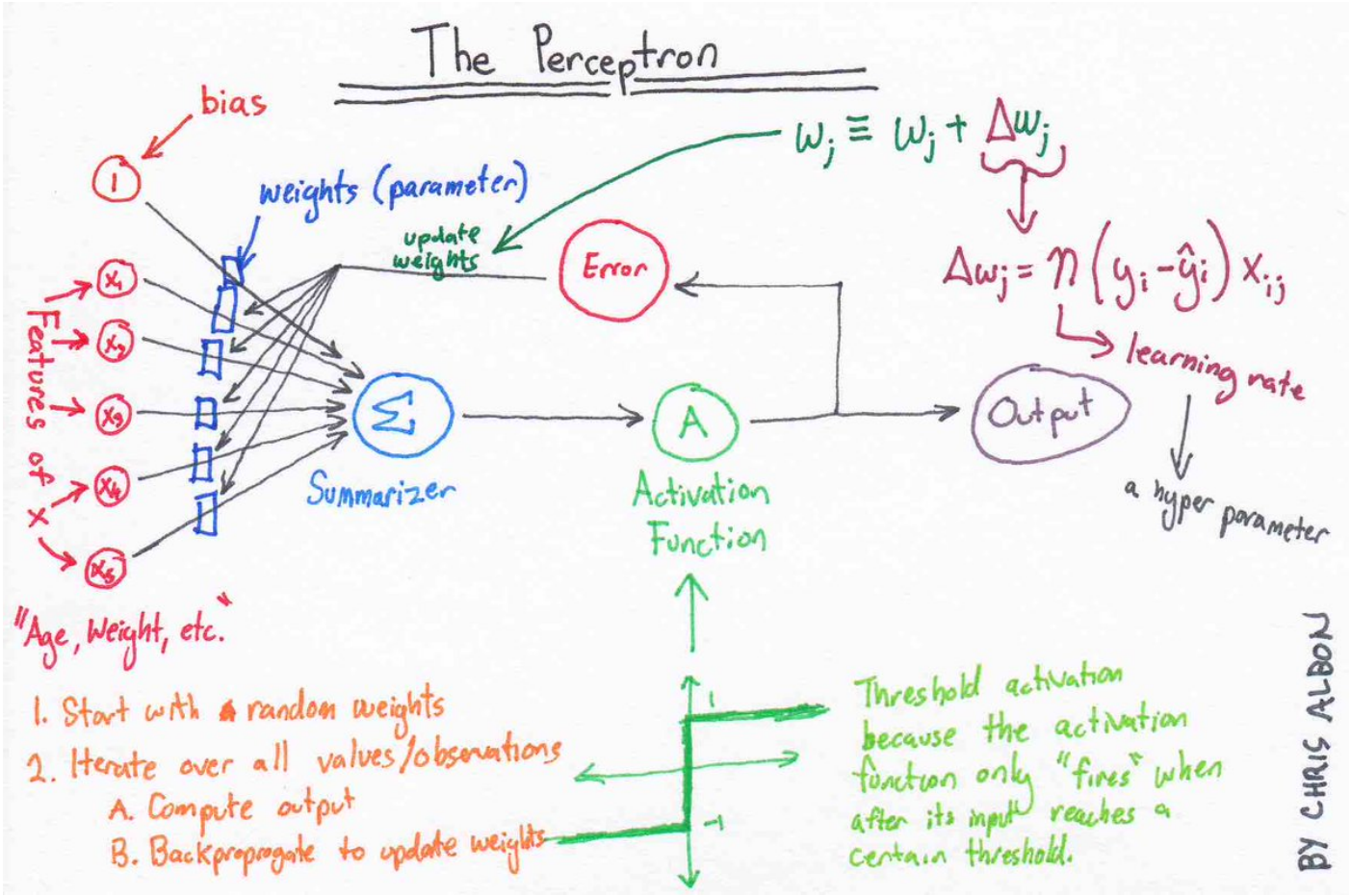
THE EFFECT OF

MODEL COMPLEXITY

TRAINING AND TEST ERROR



ChrisAlbon


The Perceptron
bias
1
weights (parameter)
update weights
$w_j \equiv w_j + \Delta w_j$
$\Delta w_j = \eta (y_i - \hat{y}_i) x_{ij}$
learning rate
a hyper parameter
Error
Features of X
x_1
x_2
x_3
x_4
x_5
"Age, Weight, etc."
Σ
Summarizer
A
Activation Function
Output
1. Start with random weights
2. Iterate over all values/observations
A. Compute output
B. Backpropogate to update weights
1
-1
Threshold activation because the activation function only "fires" when after its input reaches a certain threshold.
BY CHRIS ALBON

THE RANDOM IN RANDOM FOREST

1. Each tree gets random sample of observations with replacement.
2. Each tree gets all features, but at each node only a subset of those features are avaliable.

BY CHRIS ALBON

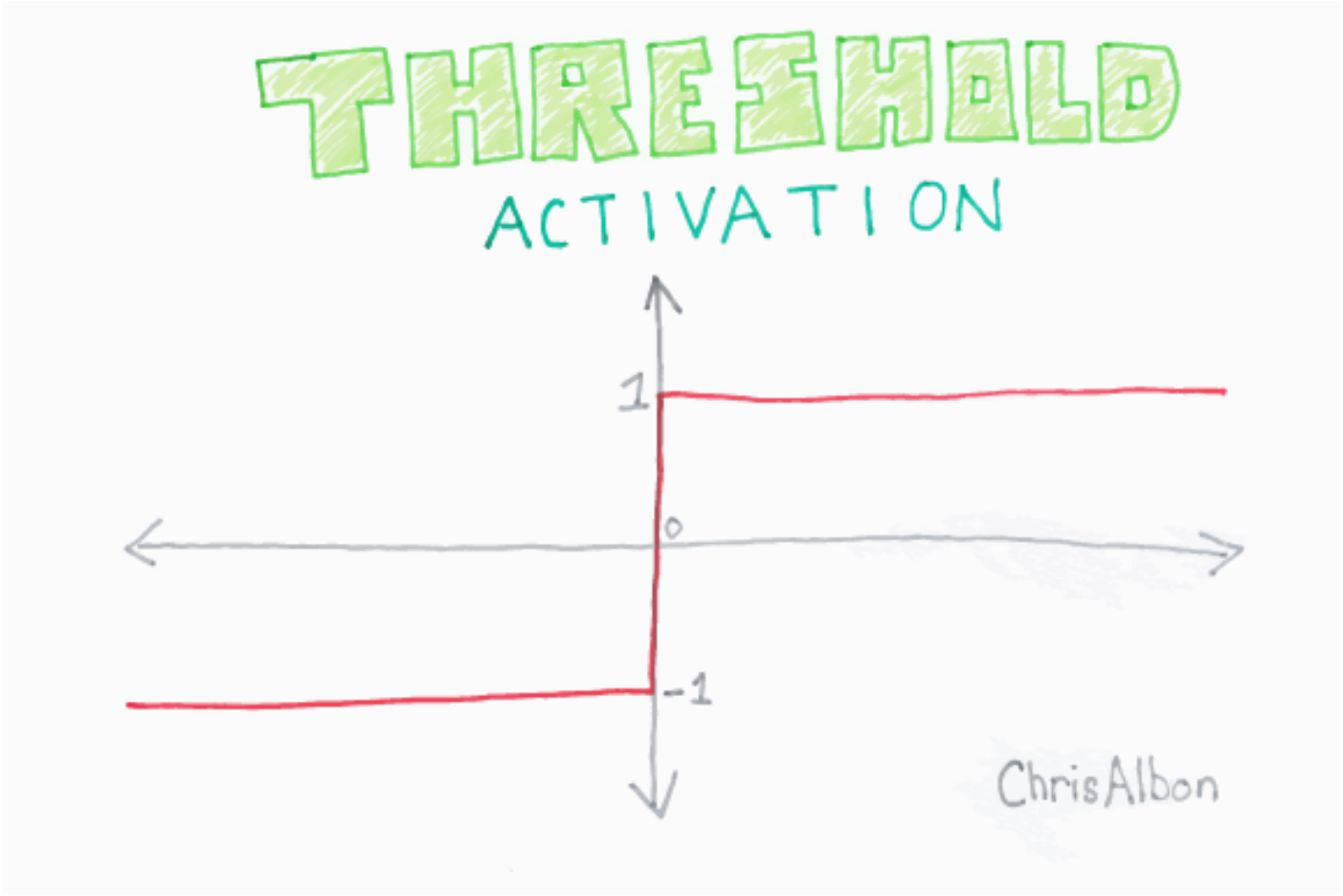
THEREFORE ∴
BECAUSE ∵

NOTATION

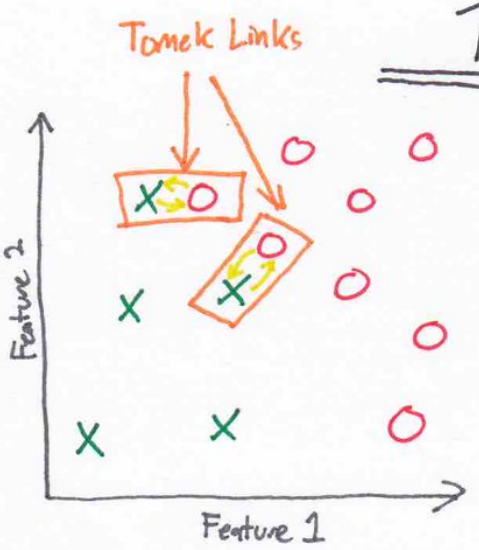
THEREFORE

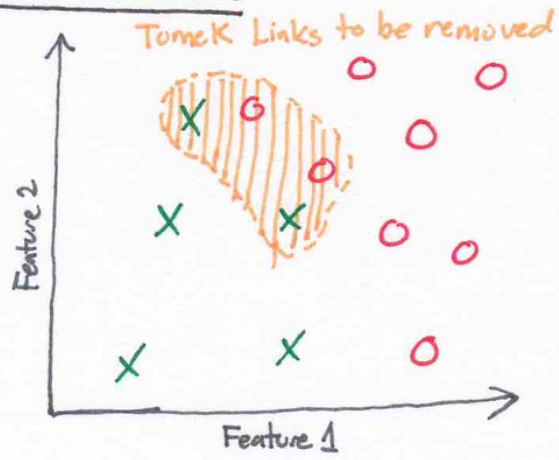
BECAUSE

BY CHRIS ALBON


THRESHOLD
ACTIVATION
1
0
-1
Chris Albon

Tomek Links





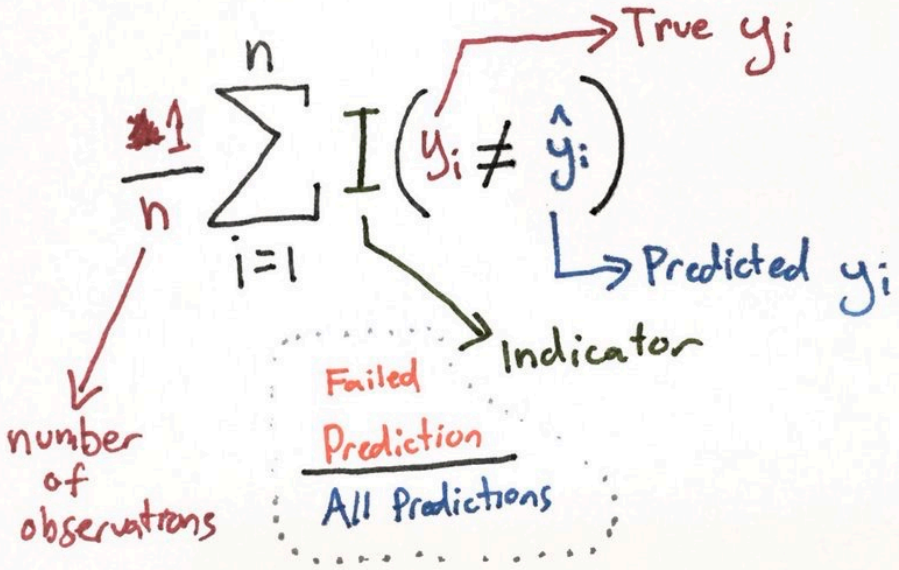
Tomek Link:

1. x's nearest neighbor is y
1. y's nearest neighbor is x
3. x and y are different classes.

Removing Tomek links has been shown to improve model performance. But problematic when n is small or classes are highly imbalanced.

BY CHRIS ALBON

Training Error Rate



Training ERROR RATE

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- n: number of observations
- I: indicator
- y_i: True y_i
- $\hat{y}_i$: Predicted y_i

$$\frac{\text{FAILED PREDICTIONS}}{\text{ALL PREDICTIONS}}$$

BY CHRIS ALBON

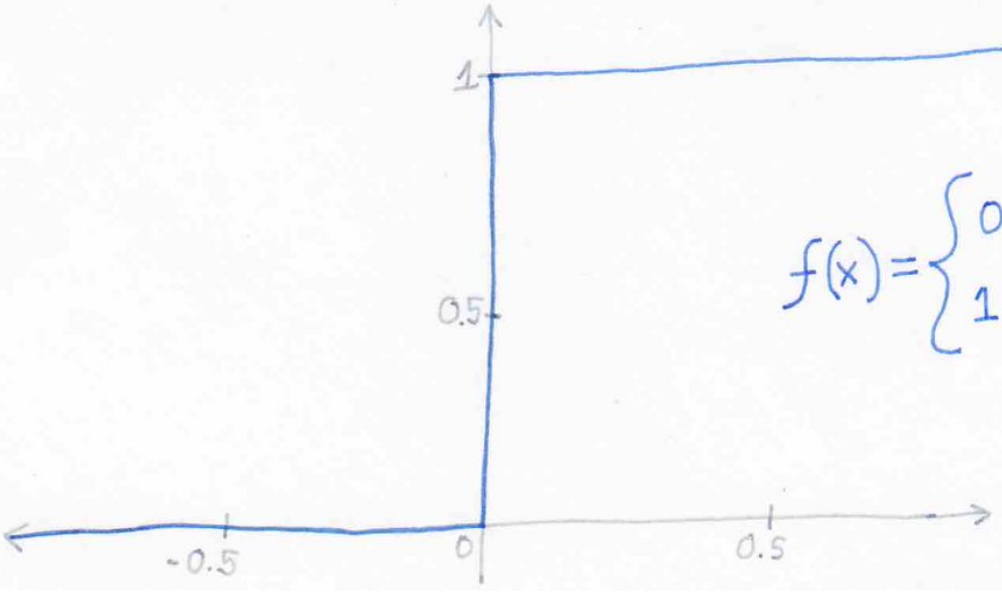
UNDERFLOW

Underflow occurs when a number is so small that it is too small to be represented by the computer. The computer will most often round these values to zero, which can be problematic because zero often behaves differently to small numbers.

Chris Albon

UNIT-STEP

ACTIVATION FUNCTION



$$f(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x \geq 0 \end{cases}$$

BY CHRIS ALBON

VANISHING GRADIENT PROBLEM

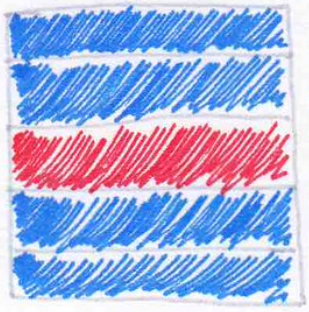
When the gradients of the loss function with respect to the parameters in the early layers of a network are very small. Causes slow learning and since many gradients are tiny, they don't contribute much to learning and can lead to poor performance

Chris Albon

VARIANCE FOR FEATURE SELECTION

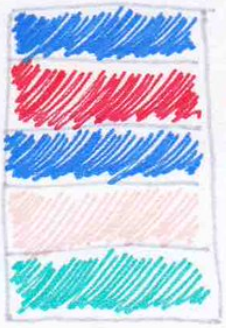
VARIANCE IS A USEFUL PROXY FOR THE INFORMATION CONTAINED IN A FEATURE. MORE VARIANCE MEANS MORE INFORMATION. AND THUS MORE USED IN TRAINING.

FEATURE 1, LOW VARIANCE



WITH LOW VARIANCE, THIS FEATURE AS LESS ABILITY TO TRAIN A MODEL.

FEATURE 2, HIGH VARIANCE



WITH HIGH VARIANCE THIS FEATURE HAS MORE INFO TO TRAIN A MODEL.

VIF

VARIANCE INFLATION FACTOR

Measures the effect of collinearity among features. Specifically, measures how much the variance of a model parameter increases if features are correlated.

To calculate VIF we make the feature the target of the model. Then run the model and calculate the R^2: $VIF_i = \frac{1}{1 - R_i^2}$

VIF = 1, not correlated. VIF > 3 correlated.

ChrisAlbon

VARIANCE

$$\mathrm{Var}\left[\hat{f}(x)\right] = E\left[\hat{f}(x)^2\right] - E\left[\hat{f}(x)\right]^2$$

Variance is the amount our predicted values would change if we had a different training dataset. It is the "flexibility" of our model, balanced against bias.

BY CHRIS ALBON

What are principal components?

PCA finds low dimensional representation of data that contains as much of the variance of the original data as possible.

↳ a measure of interestingness

Principal components are the linear combination of features that has the maximum variance out of all linear combinations that are uncorrected to the previous principal component.

Alternative interpretation: principal components are low dimensional linear surfaces closest to the observations.

Youden's J statistic

How good are we at predicting positives

How good are we at predicting negatives.

$$J = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1$$

Values range between -1 and 1. 0 = test is useless (same number of false positives as false negatives). 1 = test is perfect (no false positives and no false negatives).

Often used to select the cut-off in a ROC. Highest J is best cut-off.

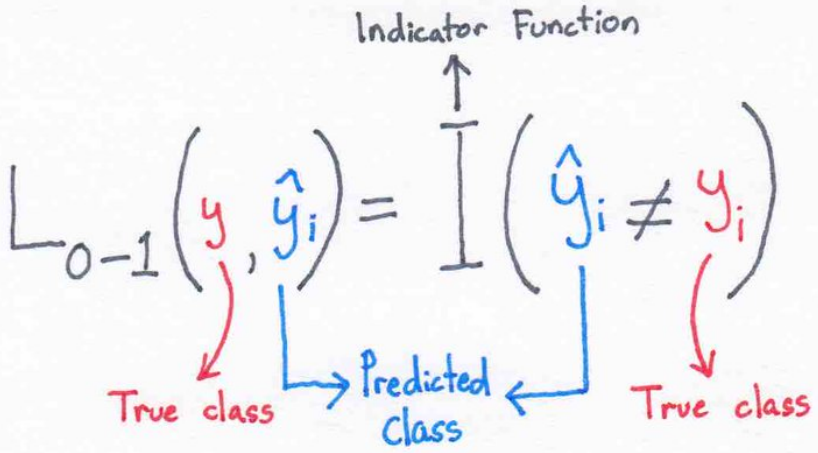
Z-score

I'm drunk!

$$\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Z-score is the number of standard deviations away from the mean.

ZERO-ONE LOSS



BY CHRIS ALBON