AdaBoost

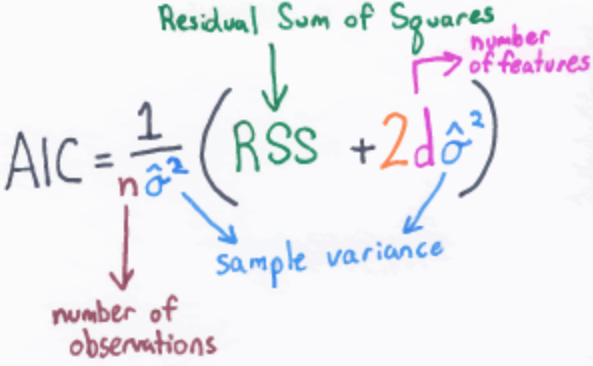
1. Assign every observation, x_i, an initial weight value, $w_i = \frac{1}{n}$, where n is the total number of observations.
2. Train a "weak" model. (most often a decision tree)
3. For each observation:
    - 3.1. If predicted incorrectly, w_i is increased
    - 3.2. If predicted correctly, w_i is decreased
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.

ChrisAlbon

AIC

"Ey-Kye-Ih-Key"

Information Criteria



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

BY CHRIS ALBON

ARCHITECTURE

OF A NEURAL NETWORK

- The architecture of a neural network refers to the units, their activation functions, how many layers etc.

- Most neural networks' architecture can be understood of as stacked layers of units.

- The best architecture for a problem should be found through experimentation using validation sets.

Chris Albon

BAGGING vs. DROPOUT

MODELS: In bagging, all models are independent.

In dropout, subnetworks share parameters.

TRAINING: In bagging, all models are trained.

In dropout, only a fraction of possible subnetworks are trained.

BAG OF WORDS

Converts text to a matrix where every row is an observation and every feature is a unique word. The value of each element in the matrix is either a binary indicator marking the presence of that word or an integer of the number of times that word appears.

Chris Albon

Bias - Variance Tradeoff

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

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

$f(x)$: true

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

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

σ_e^2: irreducible error

Bias^2

How much predicted values differ from true values.

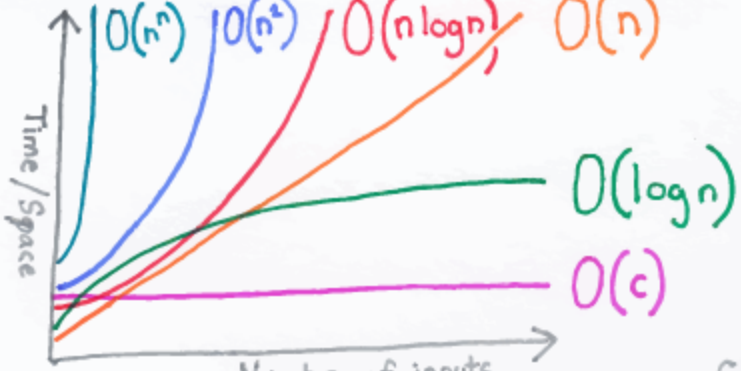
Variance

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

BY CHRIS ALBON

BIG O

Representation of an algorithm's space requirement or run time increase as the number of inputs increase



ChrisAlbon

BRIER SCORE

$$BS = \frac{1}{n} \sum_{t=1}^{n} \left(P_t - O_t \right)^2$$

Predicted Probability

number of observations

Actual outcome

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

Chris Albon

COMBINATION

Number of subsets of k items from a total of n items.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

"From n, choose k."

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.

ChrisAlbon

COST AND LOSS FUNCTIONS

Cost functions and loss functions mean the same thing. They are the objective function we are trying to train an model that minimizes.

EXAMPLE: Cross-entropy

Chris Albon

C_p

C_p is used in model selection to compare the performance of different models.

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

- n: Number of observations
- RSS: Residual Sum of squares
- d: Number of features
- $\hat{\sigma}^2$: Estimated error variance
- $2d\hat{\sigma}^2$: The penalty to adjust for the fact that training data underestimates the test error.

Chris Albon

CDF

Cumulative Distribution Function

CDF tells us the probability a random variable returns a value less than some specified value. It is the accumulation of the probability of values up to some value.

$$f(x) = P_r[X \leq y]$$

Some value $\in X$

Random variable

Chris Albon

DBSCAN

DBSCAN looks for densely packed observations and makes no assumptions about the number or shape of clusters.

1. A random observation, x_i, is selected
2. If x_i has a minimum of close neighbors, we consider it part of a cluster.
3. Step 2 is repeated recursively for all of x_i's neighbors, then neighbors' neighbors etc... These are the cluster's core members.
4. Once Step 3 runs out of observations, a new random point is chosen

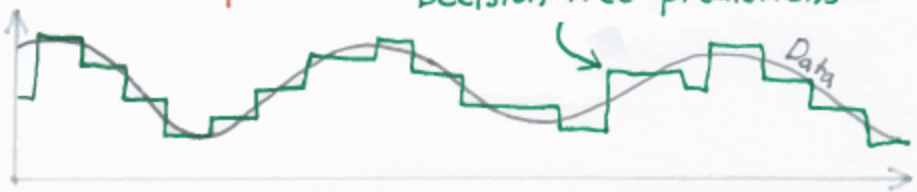
Afterwards, observations not part of a core are assigned to a nearby cluster or marked as outliers.

ChrisAlbon

DECISION TREE

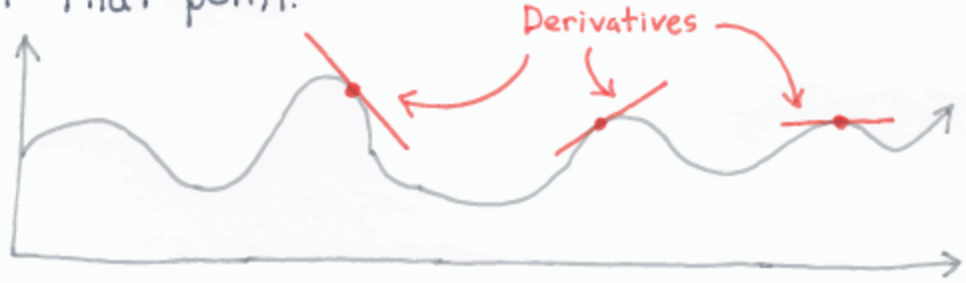
REGRESSION

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



DERIVATIVE

The derivative of a function is its rate of change. Intuitively, the first derivative at some point is the slope of the function at that point.



Chris Albon

DETERMINANTS

Determinants can be thought of as a scalar value summarizing a matrix. Formally, the determinant of a matrix is the product of all the matrix's eigen-values.

The absolute value of the determinant can be imagined as how much the matrix expands or contracts space.

Chris Albon

DOES K-NN LEARN

K-nearest neighor does not "learn" per-se. It is lazy and just memorizes the data.

Chris Albon

EARLY STOPPING

ADVANTAGES

1. Does not require altering the network architecture or training method.

2. Stops automatically at the best point rather than requiring hyperparameter tuning like weight decay.

Chris Albon

THE EFFECT OF ONE-HOT ON FEATURE IMPORTANCE

In random forests we can measure the importance of each feature. However, if our feature is a nominal categorical feature that we one-hotted, the importance of that feature will be distributed accross all of these one-hotted features.

ChrisAlbon

ELASTICNET

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

$$RSS + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2$$

Alpha

Rho

Weights

Residual sum of squares

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

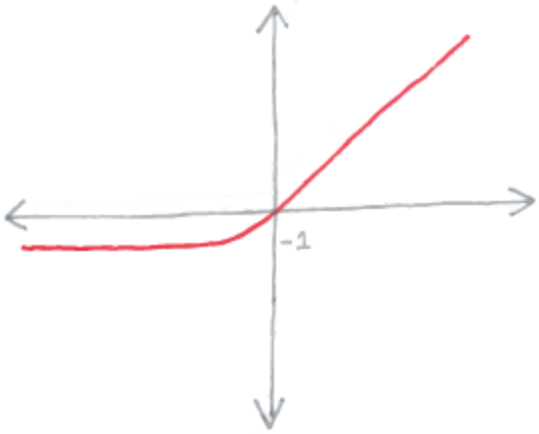
ChrisAlbon

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

EPOCH

In artifical neural networks each time all observations have been sent through the network is called an epoch. Training neural networks typically involves multiple epochs.

Chris Albon

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

ChrisAlbon

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

ChrisAlbon

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.

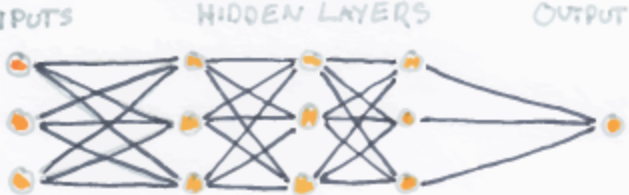
or recursive

BY CHRIS ALBON

FEEDFORWARD

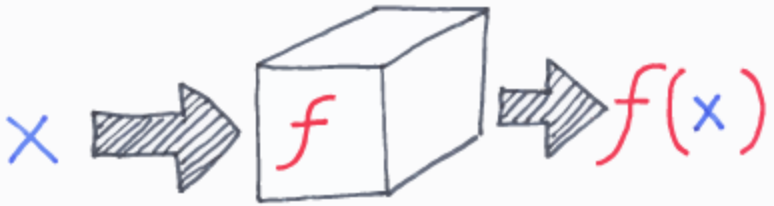
NEURAL NETWORKS

- Also called multilayer perceptron.
- Called feedforward because information moves forward from inputs to output layer.
- Structure example:



Chris Albon

FUNCTION



$f: X \to Y$

Function f maps x to y

- The domain of the function is the possible set of values of x.
- The range of the function is the possible set of values of y.

Chris Albon

GRADIENT

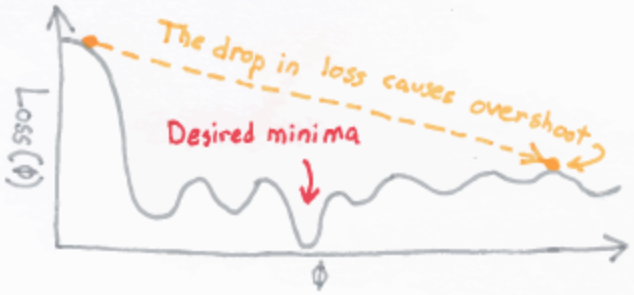
The gradient of a function, f, is a vector containing all its partial derivatives at some point, x.

$$\nabla_x f(x)$$

Chris Albon

GRADIENT CLIFF

When there is a steep drop in the loss function. Can cause a problem when the steep gradient causes the optimization algorithm to overshoot the minima.



Chris Albon

GRADIENT CLIPPING

Cliffs in the loss function can cause the learning process to overshoot a desirable minimum. This happens because the gradient at the cliff is so high. One solution is to prevent the gradient from taking extreme values:

$$\text{if } \|g\| > v: \quad g \leftarrow \frac{gv}{\|g\|}$$

gradient

Threshold

norm of gradient

ChrisAlbon

GREEK 3

LETTERS

N	ν	nu	Π	π	pi
Ξ	ξ	xi	P	ρ	rho
O	o	omicron	Σ	σ	sigma

HAMMING LOSS

Predicted class

true class

$$L_{Hamming} = \frac{\sum^{n} I\left(\hat{y}_i \neq y_i\right)}{n}$$

n ← number of observations

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

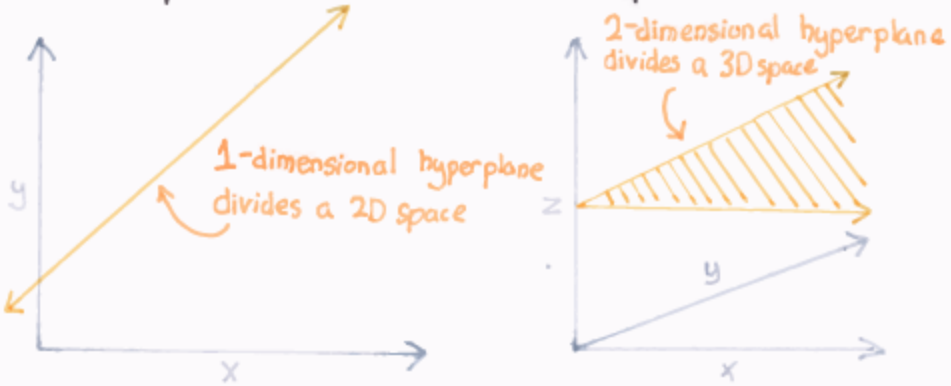
HOW NORM PENALTIES WORK

- 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

HYPERPLANE

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



ChrisAlbon

INFLECTION POINT

The point on a function where the surface changes from concave to convex or vice versa.

Chris Albon

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.

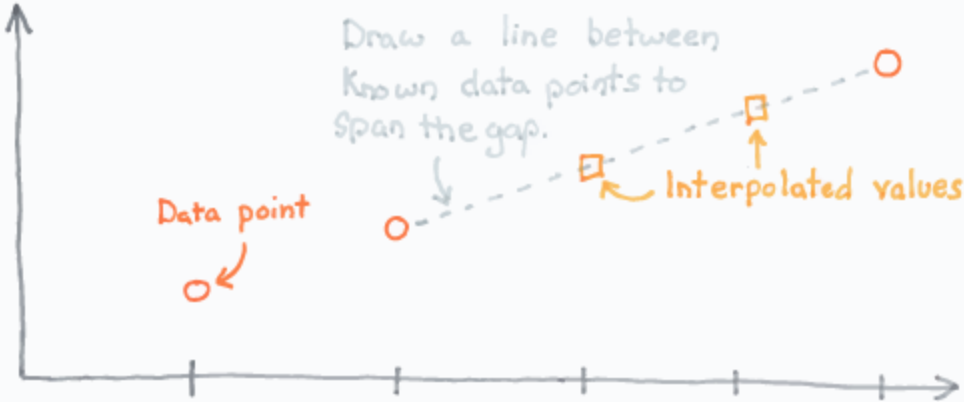
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

The interaction of features x_1 and x_2.

ChrisAlbon

INTERPOLATION

A strategy to fill gaps of missing values by drawing a line between known values.



Chris Albon

K-FOLD

CROSS-VALIDATION

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

K: # of folds

10 folds is common

$Loss_i$: For example:

- Mean squared error
- Log-loss
- Accuracy

Chris Albon

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

NEIGHBORHOOD SIZE

Small

K = Low Bias, High Variance

K = High Bias, Low Variance

LARGE

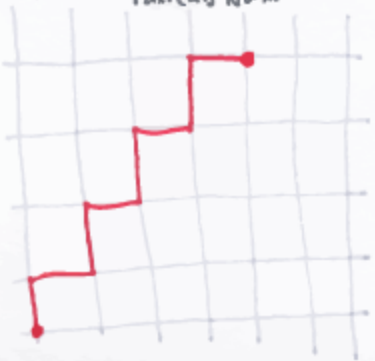
BY CHRIS ALBON

L1 Norm

(Manhattan Norm)

Also Called "Taxicab Norm"

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



ChrisAlbon

L2 NORM

(Euclidean Norm)

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

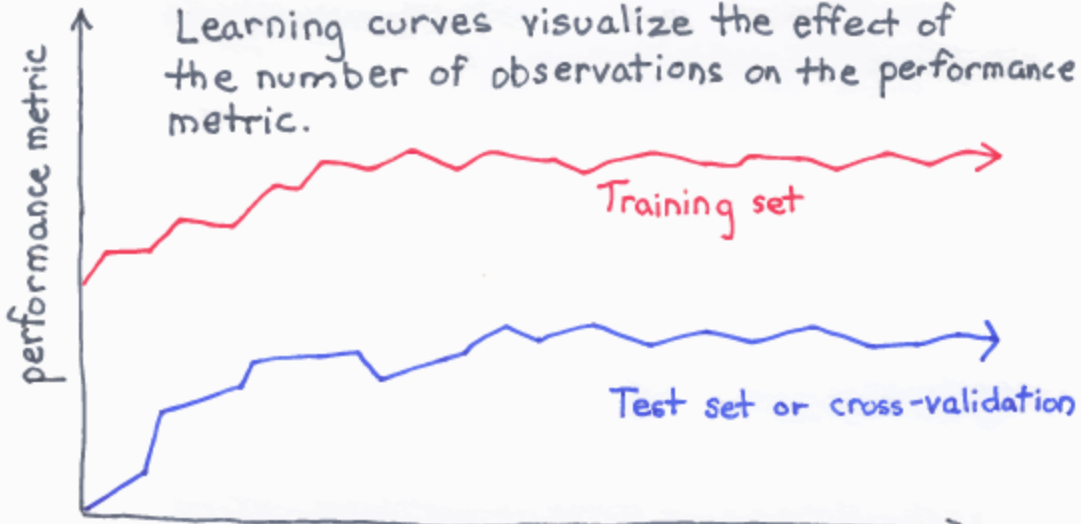
L2 norm is used in many places in machine learning such as normalizing observations and regularization.

↳ Like in NLP.

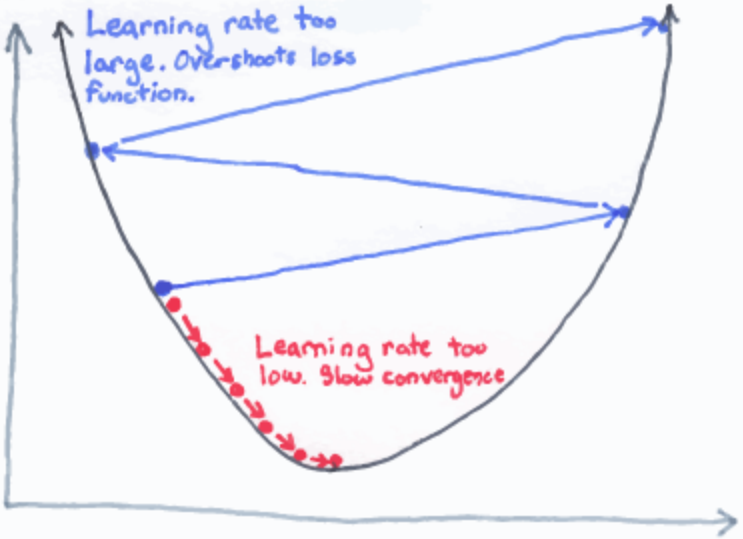
↳ Example: Ridge Regression.

Chris Albon

LEARNING CURVE



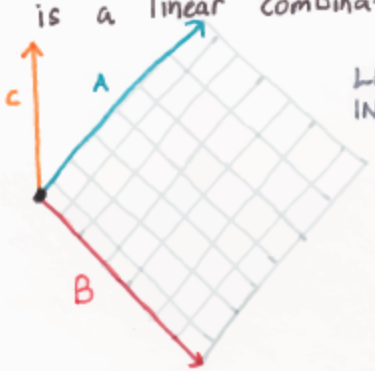
LEARNING RATE



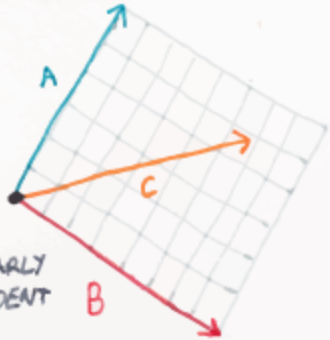
Chris Albon

LINEARLY INDEPENDENT

A matrix is linearly independent if no vector in the matrix is a linear combination of other vectors.



LINEARLY INDEPENDENT

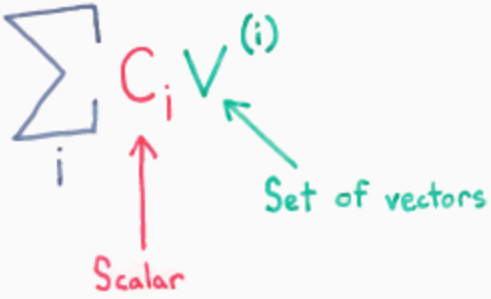


LINEARLY DEPENDENT

BY CHRIS ALBON

LINEAR COMBINATION

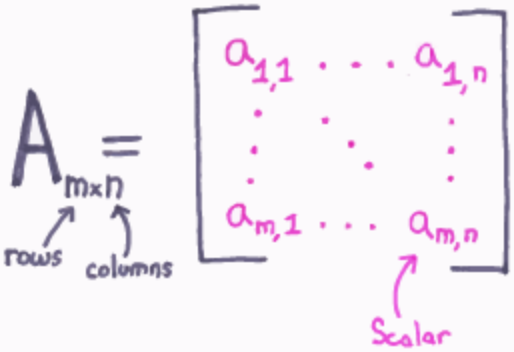
OF A SET OF VECTORS



ChrisAlbon

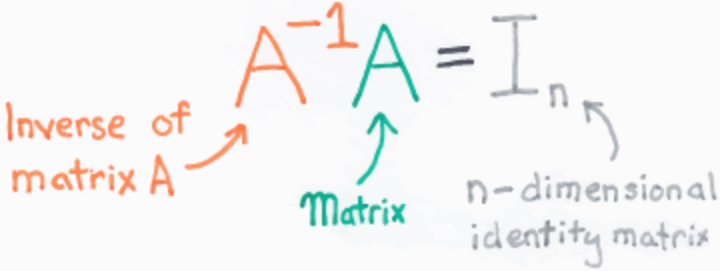
MATRICES

Two dimensional array of scalars.



ChrisAlbon

MATRIX INVERSE



ChrisAlbon

MATRIX

MULTIPLICATION

$$C_{i,j} = \sum_{k} A_{i,k} B_{k,j}$$

row index

column index

Matrix

Matrix

Matrix multiplication is associative but not commutative.

$$A(BC) = (AB)C \qquad AB \neq BA$$

Chris Albon

MATTHEWS

CORRELATION COEFFICIENT

True Positive, True Negative, False Positive, False Negative

$$M = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Ranges from -1 (perfectly wrong classifier) to 1 (perfectly right classifier).

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

MEAN ABSOLUTE ERROR

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

- n: Number of Observations
- y_i: True Target Value
- $\hat{y}_i$: Predicted Target Value

MSE

Mean Squared Error

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

true y

Predicted y

2 ← Squared

Common evaluation metric in regression

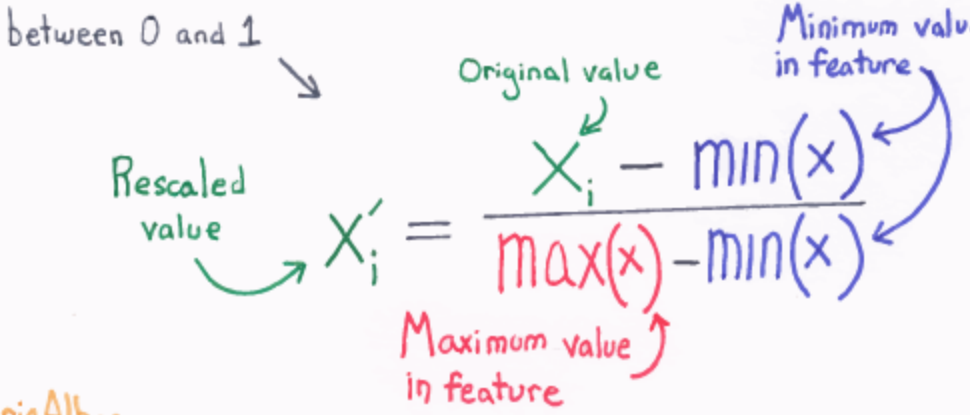
n → number of observations

Also called $\hat{y}_i$ "y hat"

BY CHRIS ALBON

MIN MAX SCALING

Rescales feature values to between 0 and 1



Chris Albon

MNAR

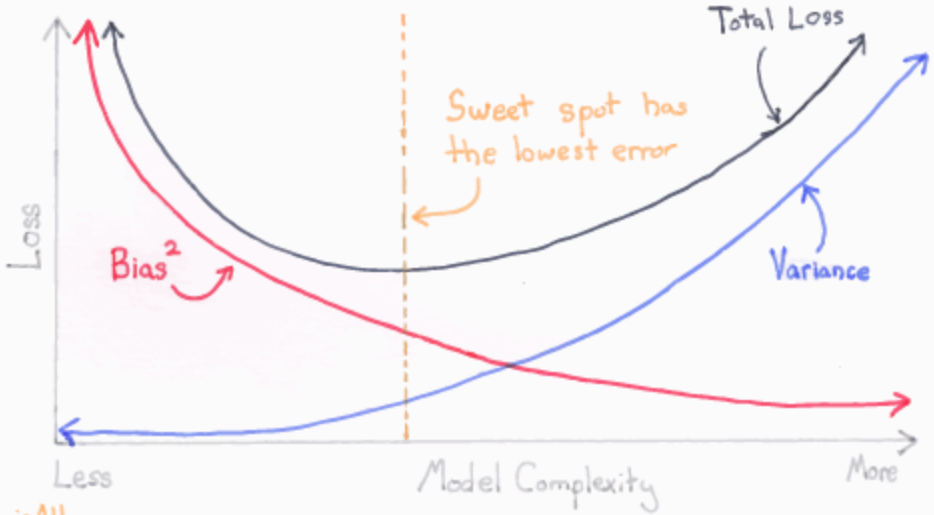
MISSING NOT AT RANDOM

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



ChrisAlbon

MODEL SELECTION

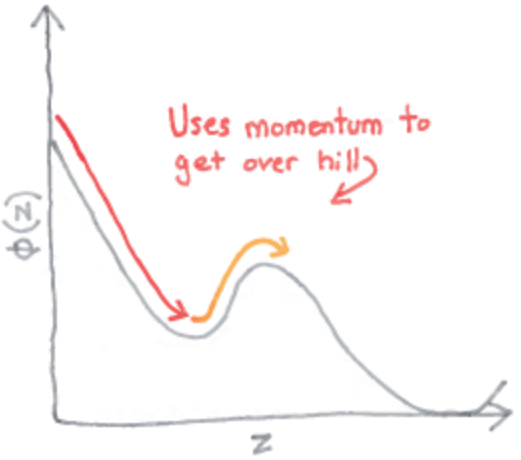
Finding the machine learning algorithm and its hyperparameter values that produce the best model.

Chris Albon

MOMENTUM

Momentum uses previous gradients to influence the movement of the optimizer.

Often used with stochastic gradient descent.



Chris Albon

MOTIVATION FOR
DEEP LEARNING

"Shallow learning" algorithms like support vector machines work well when we have well structured feature data. However, they often perform poorly in high dimensional spaces such as those found in computer vision and machine translation.

ChrisAlbon

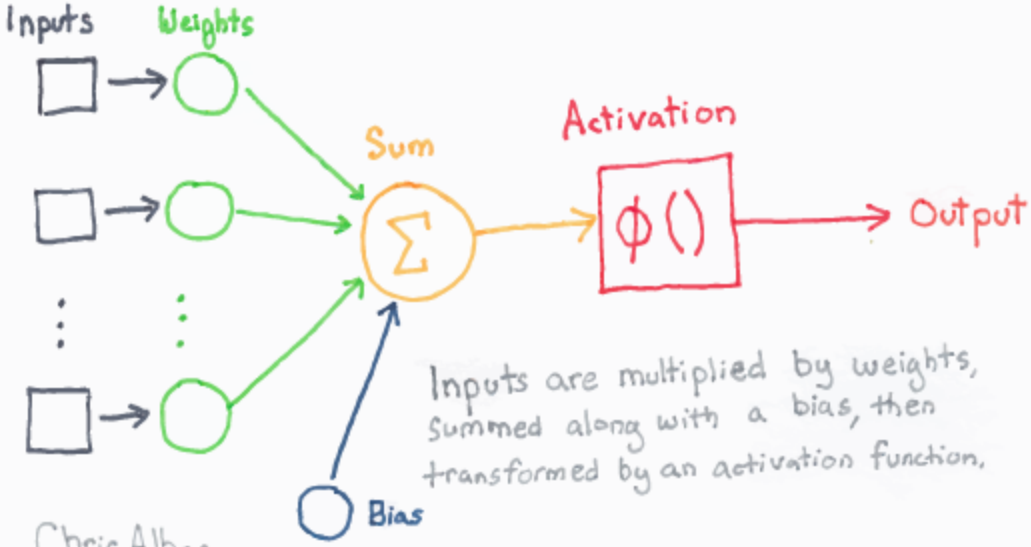
NATURAL LOG

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

Euler's number: 2.71828...

BY CHRIS ALBON

NEURON


Inputs
Weights
Sum
Activation
Output
Bias
Inputs are multiplied by weights, summed along with a bias, then transformed by an activation function.
Chris Albon

NOISY RELU

Similar to ReLU but adds Gaussian noise:

$$\phi(z) = \max(0, z + N)$$

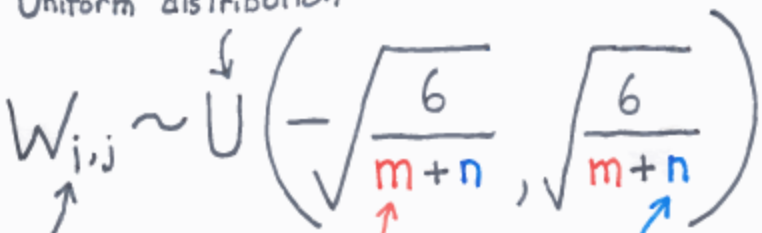
Input

Value drawn from a normal distribution:

$$\mathcal{N}(0, \sigma(z))$$

Chris Albon

NORMALIZED INITIALIZATION OF NEURAL NET PARAMETERS



Chris Albon

NORMALIZING

OBSERVATIONS

Rescaling the feature values of each observation so that they have a unit norm.

Two common norm values are L1 and L2.

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

L2 norm

ODDS RATIO

$$\frac{Pr(X_1)/Pr(\sim X_1)}{Pr(X_2)/Pr(\sim X_2)}$$

Odds of event X_1

Odds of event X_2

ChrisAlbon

OUT-OF-CORE

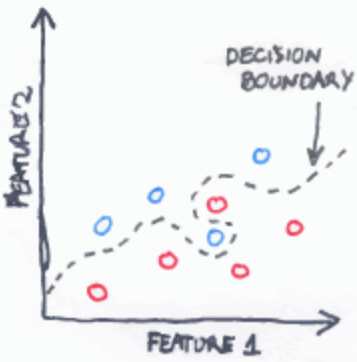
Learning methods when the data is too big to fit in a computer's RAM.

Examples of out-of-core options:

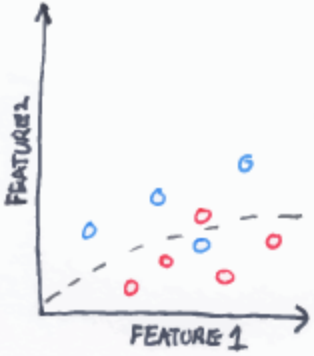
- Preprocess data in chunks
- Read and preprocess data line by line
- Incremental learning
- Stocastic techniques
- Partial fit learning methods

Chris Albon

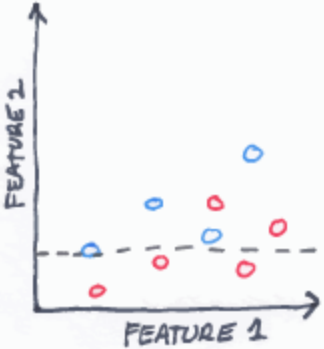
OVERFIT VS UNDERFIT



OVERFIT
"HIGH VARIANCE"



IDEAL



UNDERFIT
"HIGH BIAS"

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

PRECISION

Precision is the ability a classifier to not label a true negative observation as positive.

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

ChrisAlbon

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

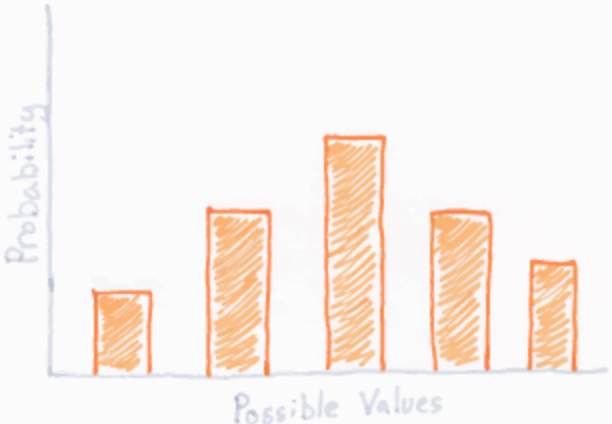
Probability Density Function

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

PMF

Probability Mass Function

The probability distribution of a discrete random variable.



ChrisAlbon

RANDOM FOREST

CLASSIFICATION

1) 1. Many trees are created using random subsets of features and bootstrapped data.

2. Each tree votes by predicting target class.

3. Votes are tallied to reach the final prediction.

CHRIS ALBON

RANDOM VARIABLE

Any variable taking on values randomly.
The values a random variable can take on is determined by a probability distribution
For example, the outcome of a dice roll is a random variable.

Chris Albon

RECALL

"Recall is about the real positives"

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

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

Chris Albon

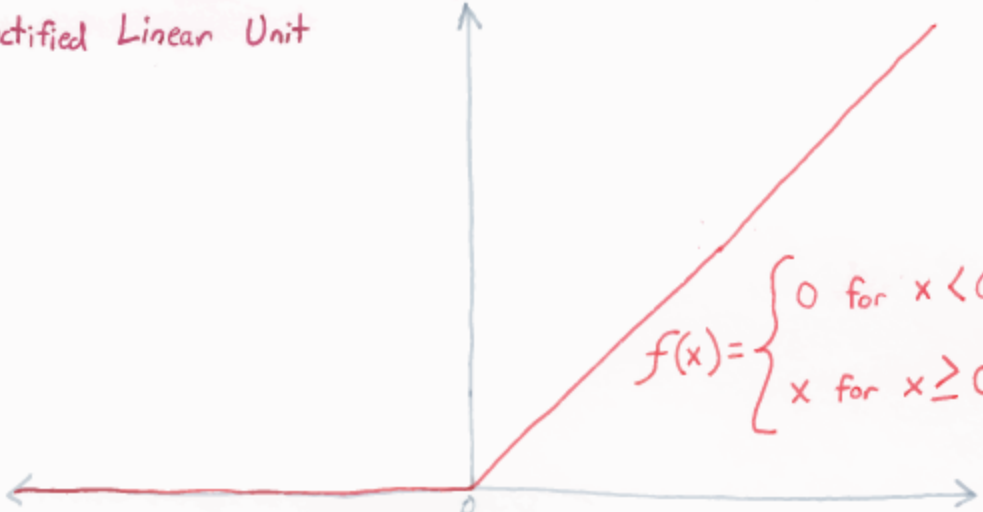
RECRESSION

Regression trains models to predict quantitative targets. Example home price.

ChrisAlbon

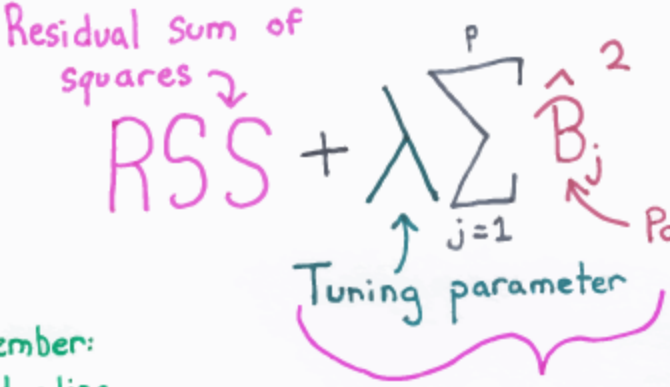
ReLU ACTIVATION FUNCTION

Rectified Linear Unit



BY CHRIS ALBON

RIDGE REGRESSION



Remember:
Standardize the data first.

Disadvantage:
parameters cannot be zero like with Lasso regression.

Chris Albon

SENSITIVITY

Also called "recall"

Everything predicted positive correctly → True Positives

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

Everything actually positive →

Chris Albon

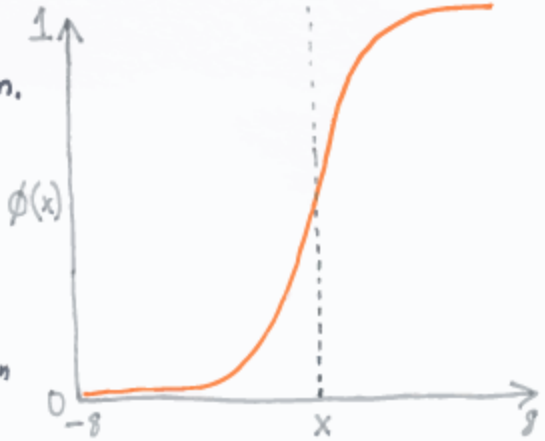
SIGMOID

ACTIVATION FUNCTION

Often used in the output unit for binary classification.

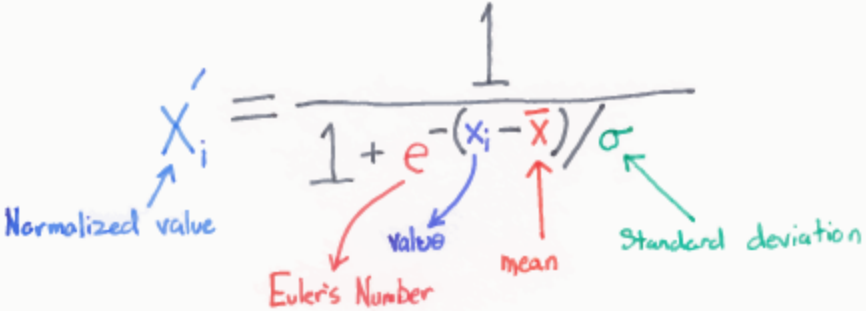
$$\phi(x) = \frac{1}{1+e^{-x}}$$

Returns a value bounded between 0 and 1.



SOFTMAX NORMALIZATION

Reduces the influence of outliers without having to drop them.



BY CHRIS ALBON

SOURCES OF UNCERTAINTY

1. Inherent randomness in the universe.
   Example: Quantum mechanics.

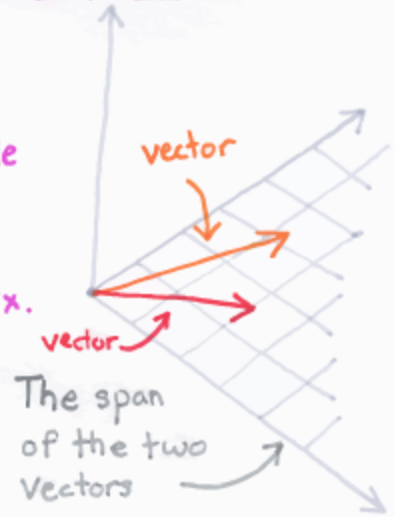
2. Inability to completely observe a phenomena, even when it is deterministic.
   Example: Observing crime.

3. Inability to perfectly model a phenomena.
   Example: Models predicting crime are simplifications.

Chris Albon

SPAN

The span of a matrix is the set of all points reachable using a linear combination of the vector of that matrix.



Chris Albon

STANDARD ERROR

OF THE MEAN

Standard error of the mean is the standard deviation of the sample mean. Estimated by dividing the Sample standard deviation by square root of the number of observations.

$$SEM = \frac{s}{\sqrt{n}}$$

s: Sample standard deviation

n: Number of observations

Chris Albon

STEMMING WORDS

Stemming reduces a word to its stem. The result is less readable by humans but makes the text more comparable across observations.

EXAMPLE: "Tradition" and "Traditional" have the same stem: "tradit"

ChrisAlbon

STOP WORDS

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

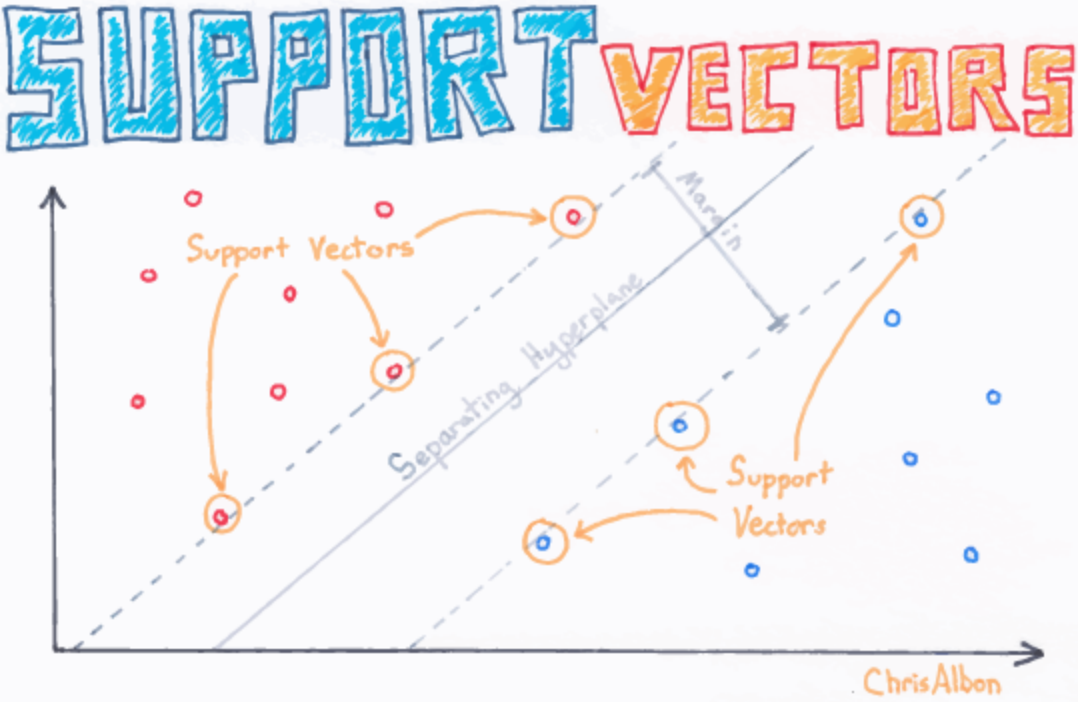
Examples
- it
- me
- myself
- we
- the
- and

Chris Albon

STRATEGIES FOR HIGHLY IMBALANCED CLASSES

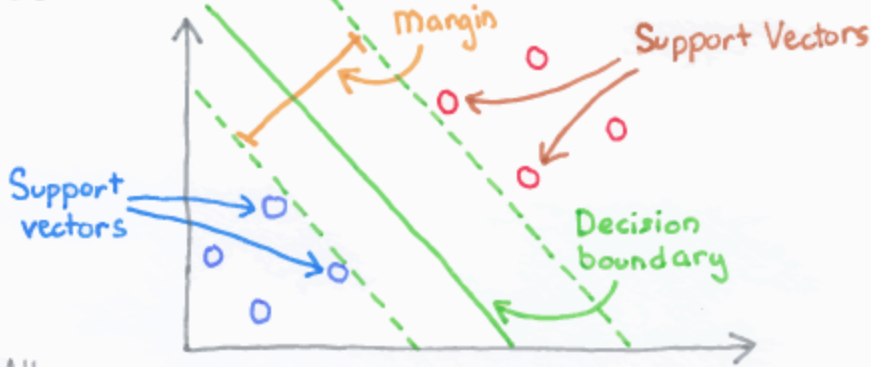
1. Collect more data
2. Choose a loss function suited for imbalanced classes like precision or recall.
3. Weight the classes
4. Downsampling and upsampling.

Chris Albon


SUPPORT VECTORS
Support Vectors
Margin
Separating Hyperplane
Support
Vectors
Chris Albon

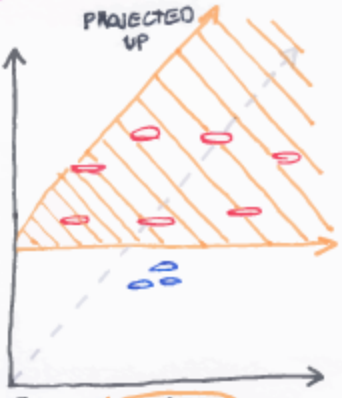
SVC

Finds the linear hyperplane that separates classes with the Maximum Margin.



Chris Albon

EYE RADIAL BASIS FUNCTION KERNEL



$$\exp\left[\frac{\|x_i - x_j\|^2}{2\sigma^2}\right]$$

Euclidean distance

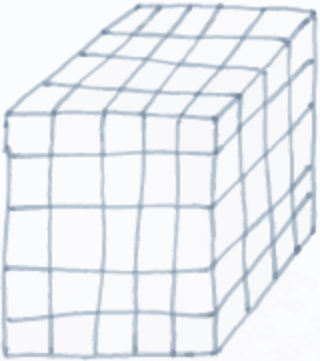
BY CHRIS ALBON

TENSORS

An n-dimensional array of numbers arranged in a "grid".

1-dimension tensor. (vector)

2-dimension tensor. (matrix)



3-dimension tensor.

Chris Albon

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

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

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

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

n ← # of documents

df(d, t) ↖ Document frequency of the term 't'

THEREFORE ∴
BECAUSE ∵

NOTATION

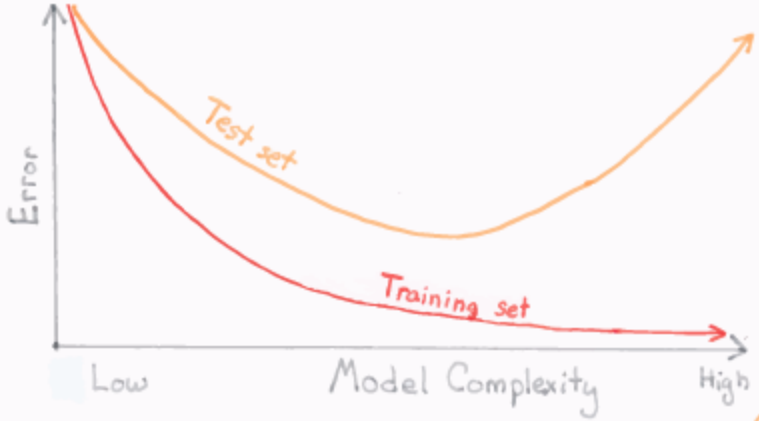
BY CHRIS ALBON

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 ON TRAINING AND TEST ERROR



ChrisAlbon

THE RANDOM IN RANDOM FOREST

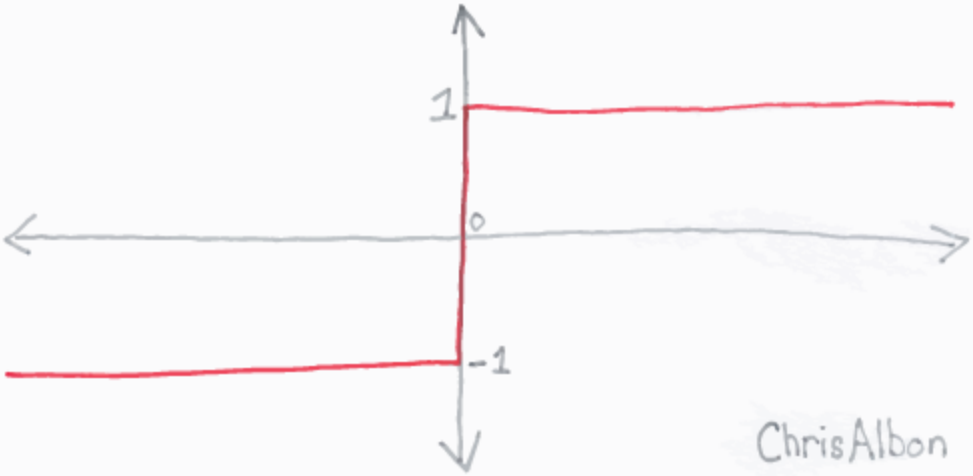
1. Each tree gets a random sample of observations with replacement.

2. Each tree gets all features, but at each node only a subset of those features are avaliable.

Chris Albon

THRESHOLD

ACTIVATION



Chris Albon

Tokenizing Text

Splitting up text into individual units like paragraphs, sentences, or words.

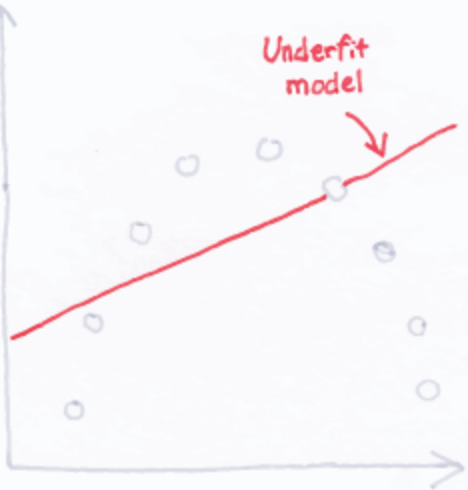
Example:

"I like birds" ⟶ "I", "like", "birds"

ChrisAlbon

UNDERFITTING

A model is underfit when it fails to capture the pattern in the data. It suffers from high bias.



ChrisAlbon

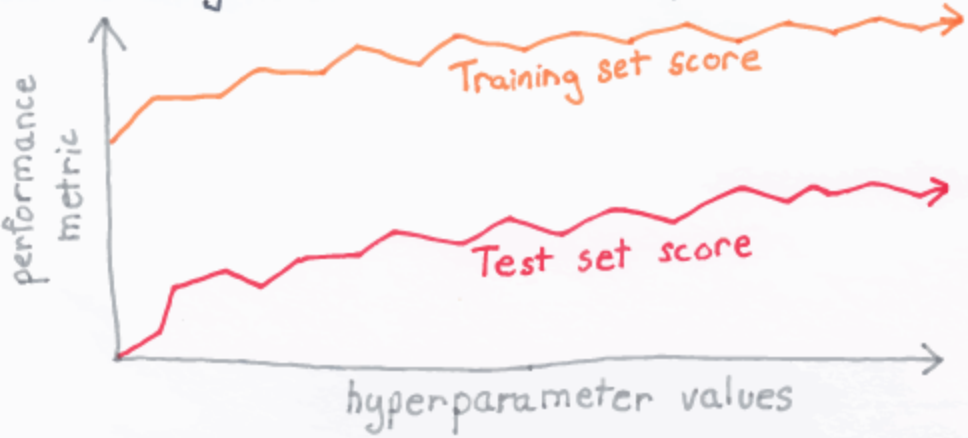
UNDERFLOW

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

Validation Curve

Validation curves visualize the performance metric over a range of values for some hyperparameter.



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

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.

Chris Albon

VECTORS

Vectors are ordered arrays of numbers (scalars).
Can be thought of as both data and geometrically:

DATA:

Scalar

$$x = [x_1\ x_2 \dots x_n]$$

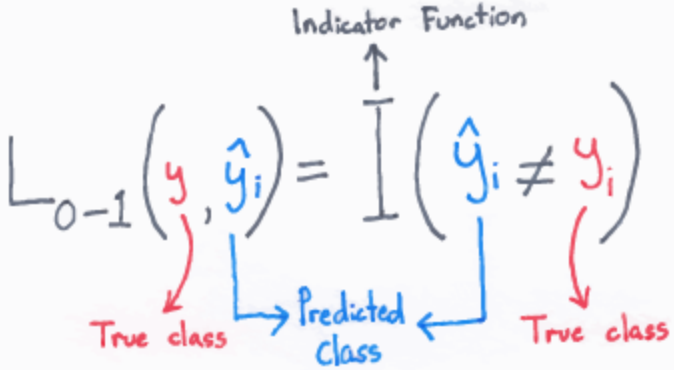
number of elements

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

GEOMETRICALLY:

BY CHRIS ALBON

ZERO-ONE LOSS



BY CHRIS ALBON