**Topic 4: Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting**

* 1. **Overfitting and its Avoidance**

A screenshot of a cell phone

Description automatically generated Hastie rental.

* + 1. **Define generalization, overfitting, fitting graph, holdout data, and base rate**

Generalization:

Generalization refers to your model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

Overfitting:

In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably". Noise.

Fitting Graph:

Curve fitting is the process of constructing a [curve](https://en.wikipedia.org/wiki/Curve), or [mathematical function](https://en.wikipedia.org/wiki/Function_(mathematics)), that has the best fit to a series of [data points](https://en.wikipedia.org/wiki/Data_points), possibly subject to constraints.

Validation Data:

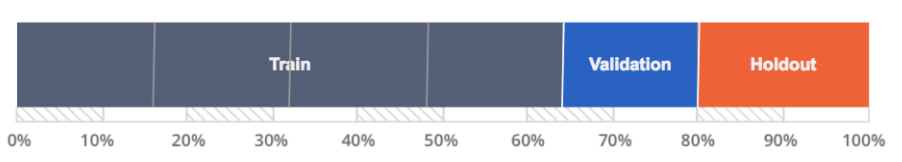
A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model’s hyperparameters.

Used to see if you are overfitting. If you train with it and show no improvement in accuracy, then you are overfitting.

Holdout Data:

This is a set of data which is held back for testing the algorithm. Sometimes referred to as “testing” data, a holdout subset provides a final estimate of the machine learning model's performance after it has been trained and validated.

Holdout sets should never be used to make decisions about which algorithms to use or for improving or tuning algorithms.



Base Rate:

In probability and statistics, base rate generally refers to the (base) class probabilities unconditioned on featural evidence, frequently also known as prior probabilities. For example, if it were the case that 1% of the public were "medical professionals", and 99% of the public were *not* "medical professionals", then the base rate of medical professionals is simply 1%.

* + 1. **Apply the concept of fitting graph to find the optimal tree induction model.**

﻿A fitting graph shows the accuracy of a model as a function of complexity. How many variables that you have used.

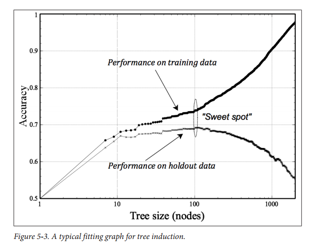
It is a way to avoid overfitting.

﻿The fitting graph (see Figure 5-1) shows the difference between a modeling procedure’s accuracy on the training data and the accuracy on holdout data as model complexity changes.

A close up of a map

Description automatically generated

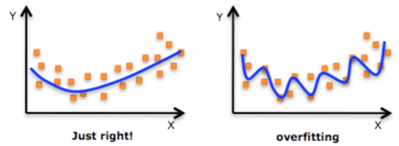
* + 1. **Define “sweet spot” for a typical fitting graph.**



Complexity vs accuracy

The sweet spot is the point just before the error on the test dataset starts to increase where the model has good skill on both the training dataset and the unseen test dataset.

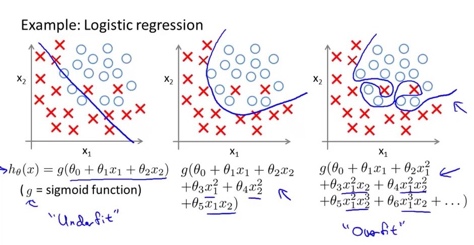
* + 1. **Apply the concept of overfitting in mathematical functions.**



In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably".

Adding too many variables to the mathematical function.

* + 1. **Analyze overfitting for logistic regression and support vector machine (SVM).**

Logistic regression overfitting. Variance is very high for an overfit logistic regression.

A screenshot of a cell phone

Description automatically generatedSupport Vector Machine overfitting

Diagram

Description automatically generated

Note: Three different models for credit analysis

* + 1. **Explain why overfitting should be of concern.**

One way to understand that intuitively is that a model may use some relevant parts of the data (signal) and some irrelevant parts (noise).

An overfitted model uses more of the noise, which increases its performance in the case of known noise (training data) and decreases its performance in the case of novel noise (test data).

* + 1. **Define cross-validation and folds.**

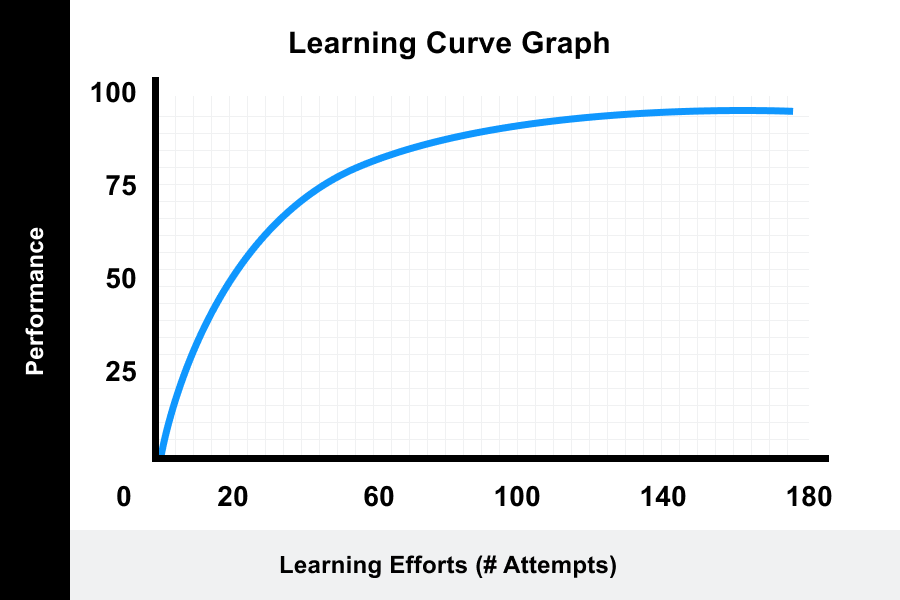
A screenshot of a computer

Description automatically generated

A fold is the section of the data.

* + 1. **Define a learning curve.**

﻿A plot of the performance VS. the amount of training data (practice) is called a learning curve. Learning effort is the training data.



* + 1. **Compare and contrast a learning curve VS. fitting graph.**

﻿ In machine learning, a learning curve (or training curve) shows the validation and training score of an estimator for varying numbers of training samples.

It is a tool to find out how much a machine learning model benefits from adding more training data and whether the estimator suffers more from a variance error or a bias error.

If both the validation score and the training score converge to a value that is too low with increasing size of the training set, it will not benefit much from more training data.

The curves will converge.

Chart, line chart

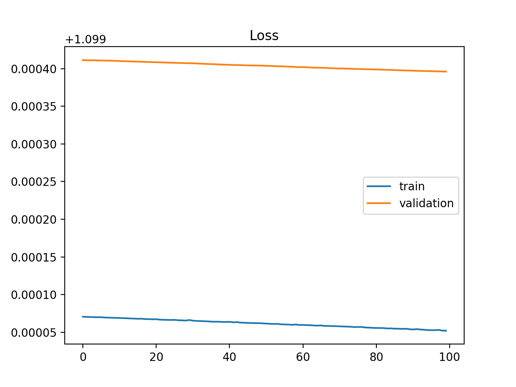
Description automatically generated

### Good Fit Learning Curves

A good fit is the goal of the learning algorithm and exists between an overfit and underfit model.

A good fit is identified by a training and validation loss that decreases to a point of stability with a minimal gap between the two final loss values.

The loss of the model will almost always be lower on the training dataset than the validation dataset. This means that we should expect some gap between the train and validation loss learning curves. This gap is referred to as the “generalization gap.”

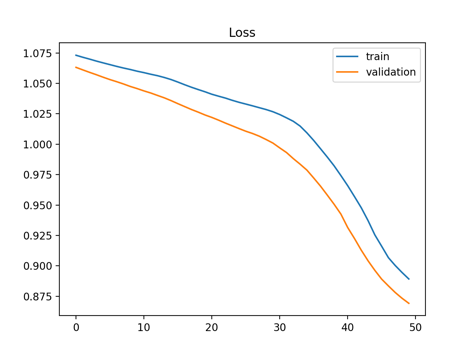


Above graph. Underfitting not showing improvement in either data set.

A plot of learning curves shows underfitting if:

The training loss remains flat regardless of training.

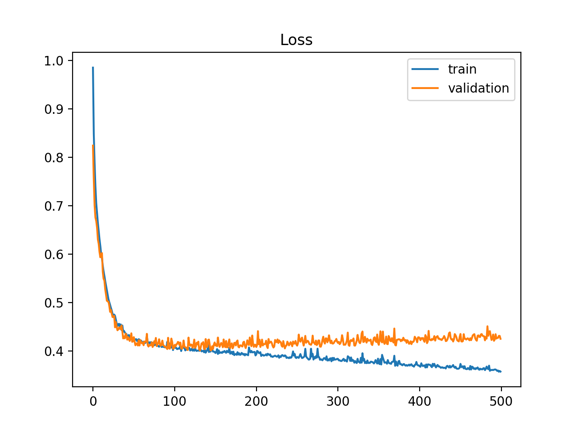
The training loss continues to decrease until the end of training. More learning to go.



Above graph. Underfitting. More learning could be achieved.

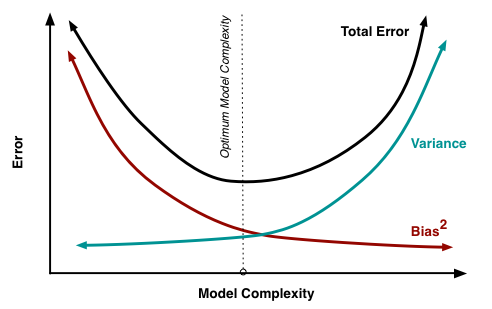
An underfit model may also be identified by a training loss that is decreasing and continues to decrease at the end of the plot.

This indicates that the model is capable of further learning and possible further improvements and that the training process was halted prematurely.



Above graph: Overfitting showing an increase in error of validation data as the while the training data error decreases. Overfit data show greater variance from the noise.

A fitting graph shows the error of training and test data plotted against model complexity of model.



Model complexity = tree with more leaves

* + 1. **Describe shape of learning curves for logistic regression and tree induction.**

Logistic will be a quick learner and then the wheels will fall off.

Chart, line chart

Description automatically generated

* + 1. **List strategies that can be used to avoid overfitting in tree induction.**

Pruning:

A bird perched on top of a grass covered field

Description automatically generated

In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus, it ends up with branches with strict rules of sparse data. This effects the accuracy when predicting samples that are not part of the training set.

* + 1. **Describe how minimum number of instances in a tree leaf can be used to limit tree size.**
* Minimum samples for a terminal node (leaf)
* Maximum depth of tree (vertical depth)
* Maximum number of terminal nodes
* Maximum features to consider for split
  + 1. **Explain how hypothesis testing can be used to limit tree induction.**

You use hypothesis testing to determine which variables are significant.

* + 1. **Define sub-training set, validation set, and nested holdout testing.**

Sub training set: To train the model

Validation set: Uses to see if you are making improvements with the training data.

Nested holdout testing: A dataset for testing.

* + 1. **Explain nested cross-validation.**

A screenshot of a cell phone

Description automatically generated

Nested for fine tuning. Validation gives you a stopping point if your model really doesn’t improve that much and you are overfitting.

* + 1. **Describe sequential forward selection and sequential backward elimination.**

**Sequential forward selection** (SFS)

An algorithm which is a bottom-up search procedure which starts from an EMPTY data set and gradually adds features selected by some evaluation function.

**Sequential backward elimination**

Backward is just the opposite. Starts with a FULL data set.

* + 1. **Describe the main idea behind regularization**

Shrinkage = Regularization

Add a small amount of bias to reduce the variance.

* + 1. **SUBSET SELECTION (Ch. 6.1)**
    2. **Define the best subset selection**

You take away one of the variables and you get a subset.

Best subsets regression is an exploratory model building regression analysis.  It compares all possible models that can be created based upon an identified set of predictors.

* + 1. **List the steps used in subset selection**

Best subsets regression is also known as “all possible regressions” and “all possible models.”

Take away variables.

Again, the name of the procedure indicates how it works. Unlike stepwise, best subsets regression fits all possible models based on the independent variables that you specify.

* + 1. **Define deviance**

A picture containing sitting, food, drawing

Description automatically generated

﻿The deviance is simply the sum of squared errors for the tree or model.

* + 1. **Define the forward stepwise selection and backward stepwise selection**

**Forward selection** is a type of stepwise regression which begins with an EMPTY model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model. Start Empty.

**Backward elimination**, which involves starting with ALL candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit. Start Full.

* + 1. **List the steps used in forward stepwise selection and backward stepwise selection.**
* Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.
* Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically insignificant loss of fit.
  + 1. **Recognize and apply the equations for Mallow’s Cp, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R2.**

LOW IS GOOD FOR ALL

**Mallow’s Cp** Image result for mallow plant images

Penalizes the number of parameters. Gaussian.

A screenshot of a cell phone

Description automatically generated

**Akaike Information Criterion (AIC)** 

**A close up of a logo

Description automatically generated**

Penalizes the number of parameters.

For a simpler model.

Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

**Bayesian Information Criterion (BIC)** A drawing of a cartoon character

Description automatically generated

Penalizes the number of parameters.

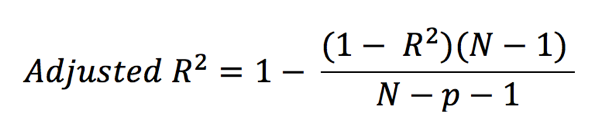
Similar to AIC but it penalizes the number of parameters more than AIC.

BIC is harsher. Note: think of the harsh Baldwin pen stab in bond pit.

A screenshot of a cell phone

Description automatically generated

**Adjusted R-squared**



Additional info:

LOW IS GOOD

**Mallow’s Cp** 

Addresses the issue of overfitting. It penalizes the complexity or overuse of parameters in the model. A small value of Cp means that the model is relatively precise.

It is the same as the Akaike information criterion but in the special case of Gaussian linear regression.

**Akaike Information Criterion (AIC)** 

Note: Keep AIC low. Get the extra parameters to avoid overfitting. AIC penalizes extra parameters.

The Akaike Information Criterion (AIC) is an estimator of the relative Quality of statistical models for a given set of data.

A person in a suit and tie

Description automatically generated with medium confidence Note: Hirotugu Akaike

AIC is now widely used for [model selection](https://en.wikipedia.org/wiki/Model_selection), which is commonly the most difficult aspect of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference); additionally, AIC is the basis of a paradigm for the [foundations of statistics](https://en.wikipedia.org/wiki/Foundations_of_statistics).

Akaike also made major contributions to the study of [time series](https://en.wikipedia.org/wiki/Time_series). As well, he had a large role in the general development of statistics in Japan.

Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

**Bayesian Information Criterion (BIC)** A drawing of a cartoon character

Description automatically generated

In statistics, the Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the LOWEST BIC is preferred.

It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. More severe. Think of Baldwin stabbing with a BIC pen.

**Adjusted R-squared.** 

If you add complete rubbish variables, then your adjusted R-squared will go down. It takes into account rubbish variables.

* + 1. **Shrinkage methods (Shrinkage = Regularizing. Ridge Regression and Lasso)**
  1. **Define Ridge Regression, tuning parameter (think tuning fork which is lambda), and shrinkage penalty**

Shrinkage = Regularization

This **shrinkage** (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.

Classification and Regression are Supervised learning. Clustering is unsupervised.

In Ridge Regression, you add a bias to the original fit. This gives you better long-term predictions.

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimate (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.

By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

* 1. **Define L2 norm and scale equivalent**

L2 Regularization or Ridge Regression or L2 norm. Note: Think 2 R’s for R-squared.

MSE is L2. Uses squared and will be more influenced by outliers.

Euclidean.

The L2 norm calculates the distance of the vector coordinate from the origin of the vector space. As such, it is also known as the Euclidean norm as it is calculated as the Euclidean distance from the origin. The result is a positive distance value.

A close up of a map

Description automatically generated

* 1. **Define standardizing the predictors**

Standardization of the predictors. Put them on the same scale.

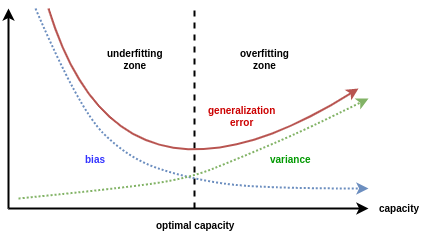
In regression analysis, you need to standardize the independent variables when your model contains polynomial terms to model curvature or interaction terms.

In statistics, standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables.

Typically, to standardize variables, you calculate the mean and standard deviation for a variable.

Then, for each observed value of the variable, you subtract the mean and divide by the standard deviation.

* 1. **Describe bias-variance trade off**

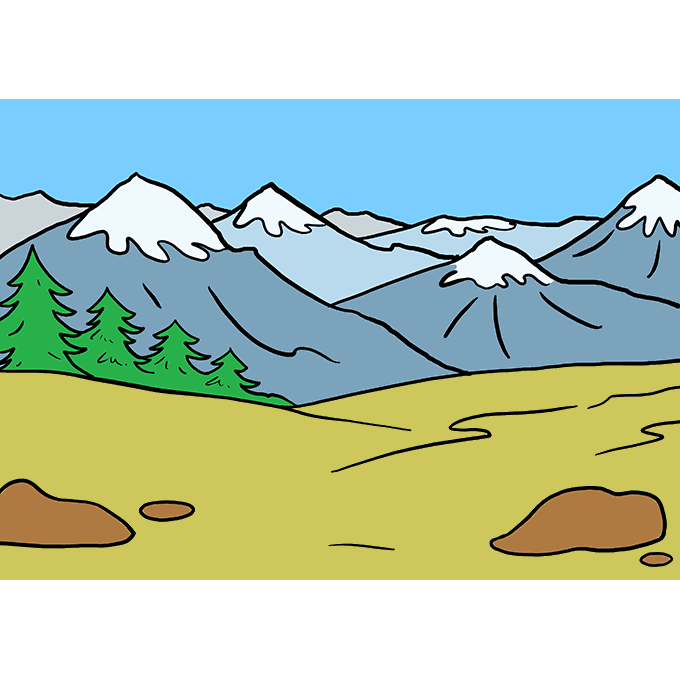


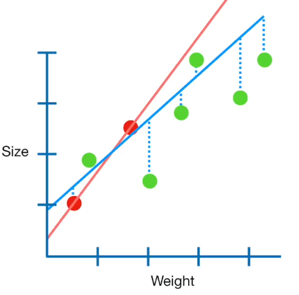
In supervised machine learning an algorithm learns a model from training data.

The goal of any supervised machine learning algorithm is to best estimate the mapping function (f) for the output variable (Y) given the input data (X). The mapping function is often called the target function because it is the function that a given supervised machine learning algorithm aims to approximate.

The prediction error for any machine learning algorithm can be broken down into three parts:

* Bias Error
* Variance Error
* Irreducible Error
  1. **Describe the Ridge Regression**





A picture containing knife, table

Description automatically generated

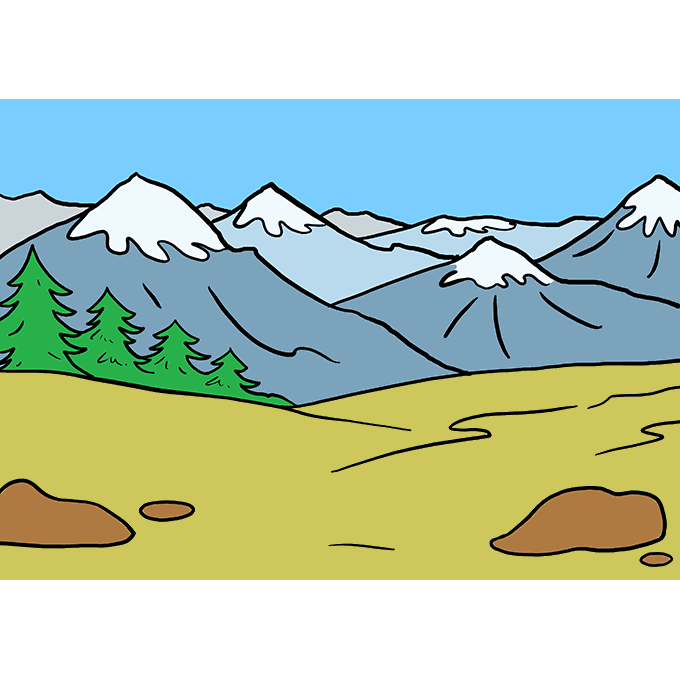
Note: Ridge Regression has two R’s for Squared (the slope). And L2 for 2 r’s.

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated).

In multicollinearity, even though the least squares estimate (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.

By adding a degree of BIAS (title to linear regression line) to the regression estimates, ridge regression reduces the standard errors.

* 1. **Describe how the Ridge Regression improves upon least squares**



By adding a degree of BIAS (known as the penalty) to the regression estimates, ridge regression reduces the standard errors. Better predictor.

* 1. **Describe the advantage of LASSO over the Ridge Regression**

Lasso in the weak variables.

****

**Minimize 🡪**

A picture containing knife, table

Description automatically generated

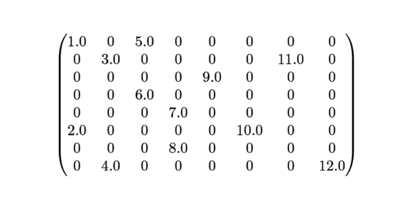
Note: **L**asso has LINES for absolute value.

The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). LASSO in the useless variables.

Advantage: It can reduce a massive model to just the most important variables.

LASSO can have a slope of zero.

* 1. **Define a sparse model**



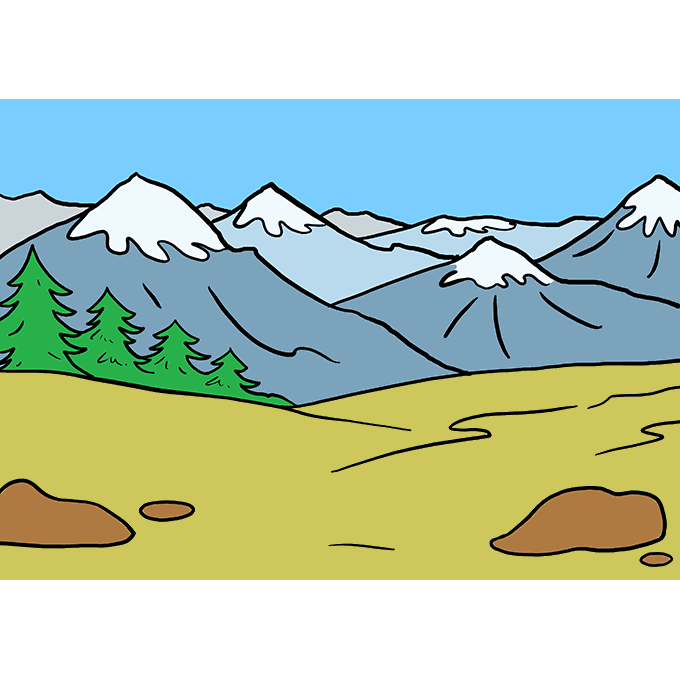
Matrices that contain mostly zero values are called sparse, distinct from matrices where most of the values are non-zero, called dense.

* 1. **Describe the variable selection property of the LASSO.**

In statistics and machine learning, **LASSO** (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

* 1. **Compare the LASSO to the Ridge Regression.**

Lasso regression gets rid of the useless variables.

**** VS. 

Lasso “catches” the worthless variables.

Lasso good when model has unnecessary variable.

Ridge Regression is good when all of the variables are useful.

* 1. **Describe how to select the tuning parameter (λ)**

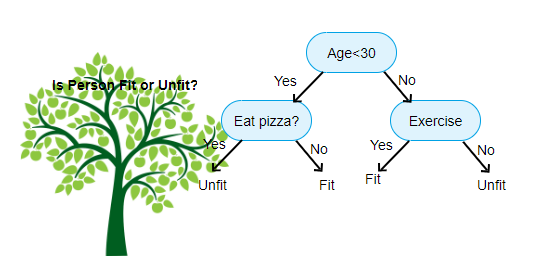
Note: Think lambda looks like a tuning (fork) parameter.

A **tuning parameter**(λ), sometimes called a penalty parameter, controls the strength of the penalty term in Ridge Regression and LASSO regression.

It is basically the amount of shrinkage, where data values are shrunk towards a central point, like the mean.

Shrinkage results in **simple, sparse models which are easier to analyze than high-dimensional data models with large numbers of parameters.**

* **L1 regularization** adds an L1 penalty equal to the absolute value of the magnitude of coefficients. In other words, it limits the size of the coefficients. L1 can yield sparse models (i.e. models with few coefficients); Some coefficients can become zero and eliminated. Lasso regression uses this method.
* **L2 regularization**adds an L2 penalty equal to the square of the magnitude of coefficients. L2 will not yield sparse models and all coefficients are shrunk by the same factor (none are eliminated). Ridge regression and SVMs use this method. R2
* **Elastic nets** combine L1 & L2 methods, but do add a hyperparameter
  + 1. **Tree Based methods**
  1. **Interpret as well as predict using a given decision tree**



* 1. **Describe the advantages and disadvantages of decision trees compared to other classification and regression methods.**

Advantages:

* Are simple to understand and interpret.
* Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
* Help determine worst, best and expected values for different scenarios.
* Use a [white box](https://en.wikipedia.org/wiki/White_box_(software_engineering)) model. If a given result is provided by a model.
* Can be combined with other decision techniques.
* In general, decision tree classifiers have good accuracy.
* The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.
* Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans.
* The learning and classification steps of decision tree induction are simple and fast.

Disadvantages of decision trees:

* They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
* They are often relatively inaccurate. Many other predictors perform better with similar data. This can be remedied by replacing a single decision tree with a [random forest](https://en.wikipedia.org/wiki/Random_forest) of decision trees, but a random forest is not as easy to interpret as a single decision tree.
* For data including categorical variables with different number of levels, information gain in decision trees is biased in favor of those attributes with more levels.
* Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.
  1. **Describe recursive binary splitting for constructing regression trees.**

A greedy approach is used to divide the space called recursive binary splitting.

Droste effect. Picture appearing within itself.

In a less specific sense, *binary splitting* may also refer to any divide and conquer algorithm that always divides the problem in two halves.

This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function.

The split with the best cost (lowest cost because we minimize cost) is selected.

* 1. **Describe tree pruning, specifically cost complexity (weakest link) pruning.**

Pruning a tree prevents overfitting to the training data.

* 1. **Describe the construction of classification trees using classification error rate, Gini-index, and cross entropy.**

Classification error rate

Everything False/ Everything.

Complement of Accuracy (1-Accuracy).

Gini-index

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

Purity measure.

Cross entropy

Cross-entropy is a measure of the difference between two probability distributions for a given random variable or set of events.

Cross-entropy is commonly used in machine learning as a loss function.

* 1. **Contrast tree-based methods VS linear models.**

Tree based model does not require linear data.

* 1. **Describe bagging and out-of-bag error estimation.**

Bagging = Bootstrap Aggregation

Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement.

Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.

* 1. Describe Random Forests.

Random Forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees.



* 1. **Describe boosting as an approach for improving the prediction results from decision trees.**

Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors.

In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.