Machine Learning Basics

Machine Learning in Python Workshop
DLab
UC Berkeley

What is Machine Learning?

“Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data [...] in order to make predictions or decisions without being explicitly programmed to perform the task”

- Wikipedia

"A computer program is said to learn from experience E, with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

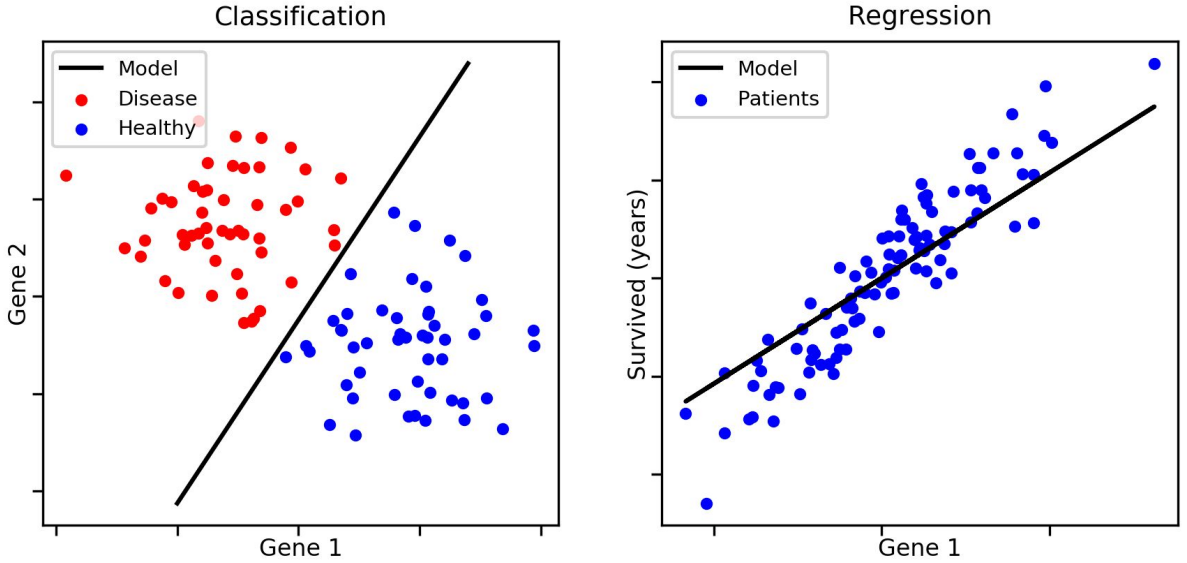
- Tom Mitchell, (1997). Machine Learning. McGraw Hill.

Types of Machine Learning Algorithms

- **Supervised Learning**: Learning a relationship between input variables, known as features, and an output variable. You can think of this as learning a (possibly very complicated!) function that maps from features to outputs.
  - Features are also known as independent variables, regressors, predictor variables, covariates
  - Outputs are also known as dependent variable, outcome variable, response variable, etc.
- **Unsupervised Learning**: Learning patterns or structure across a set of variables by finding statistical associations between those variables. Uses include: discovering latent clusters or dimensions, reducing dimensionality, outlier detection, probability density estimation.
- **Reinforcement Learning**: Allows computers to learn to accomplish a task without an explicit dataset, rather using only a “reward function” indicating what is closer or farther from the task. (Think the childs game hotter-colder).

Supervised Learning: Classification & Regression

- **Classification**: When the output variable is discrete (e.g categorical), such as a simple yes/no, or one of N mutually exclusive categories.
- **Regression**: When the output variable is a continuous number.

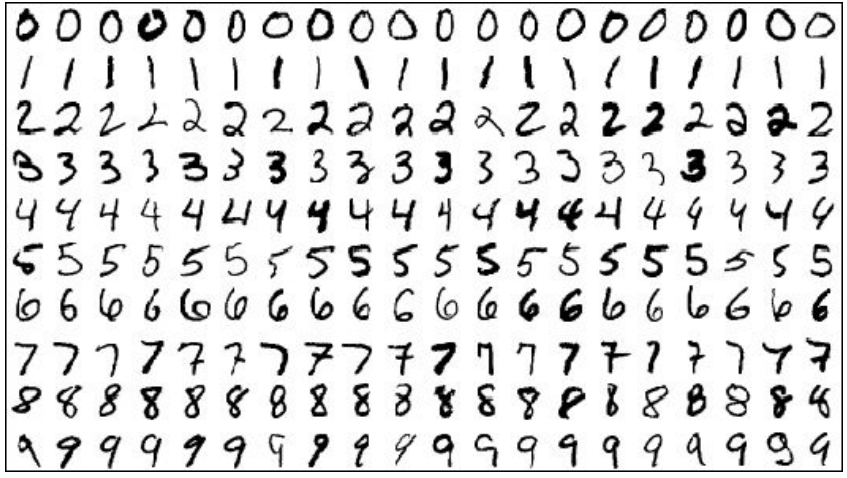


Supervised Learning Algorithms

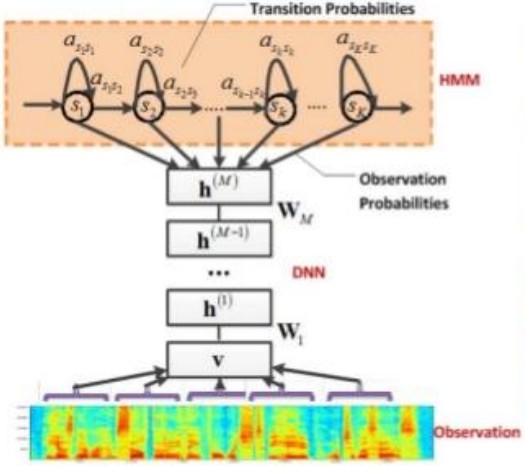
1. Generalized Linear Model
2. K-Nearest-Neighbors
3. Support Vector Machines (SVM)
4. Deep Artificial Neural Networks
5. Decision Trees & Random Forests
6. + many more!!!

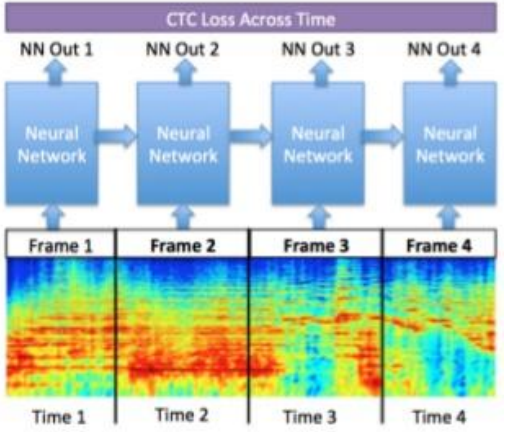
Supervised Learning Examples

Handwritten Digit Recognition



Speech Recognition





3D Face Modeling
Nine Layer-Deep Neural Network

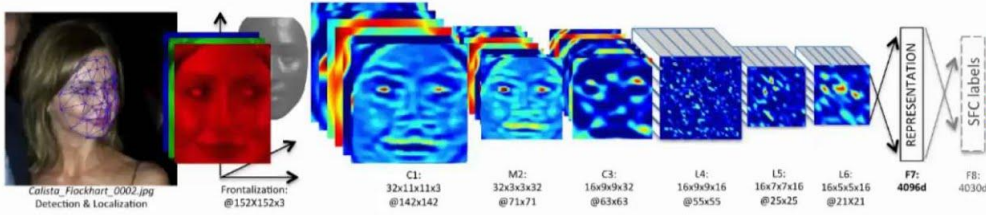


Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

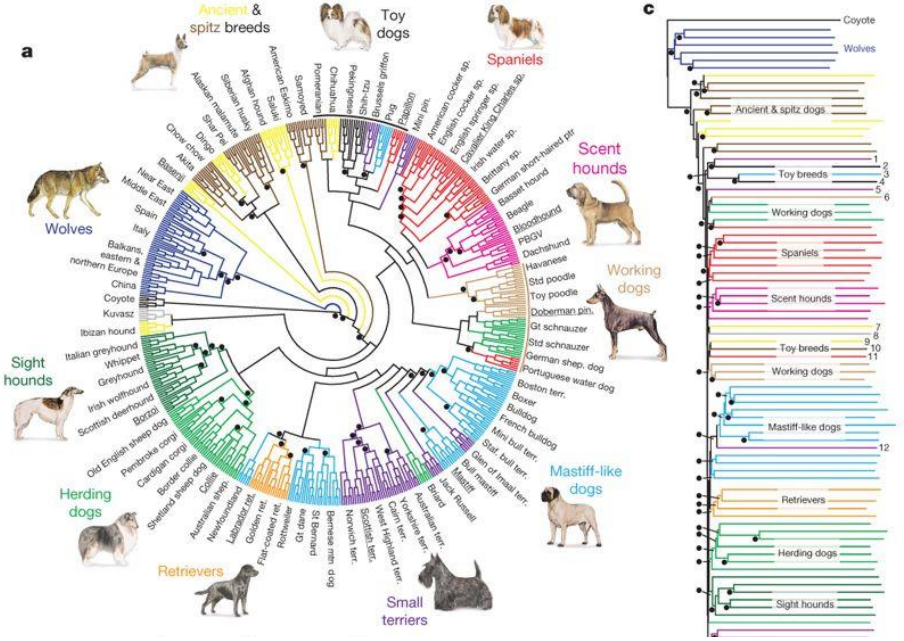
Face Recognition: Detect → Alignment → Represent → Classify

Unsupervised Learning Algorithms

1. Principal Components Analysis (PCA)
2. Independent Components Analysis (ICA)
3. Multi-Dimensional Scaling (MDS)
4. K-means clustering
5. Hierarchical Clustering
6. Gaussian Mixture Models
7. + many more!!!

Unsupervised Learning Examples

Hierarchical Clustering in Genomics



Cocktail Party Problem (ICA)

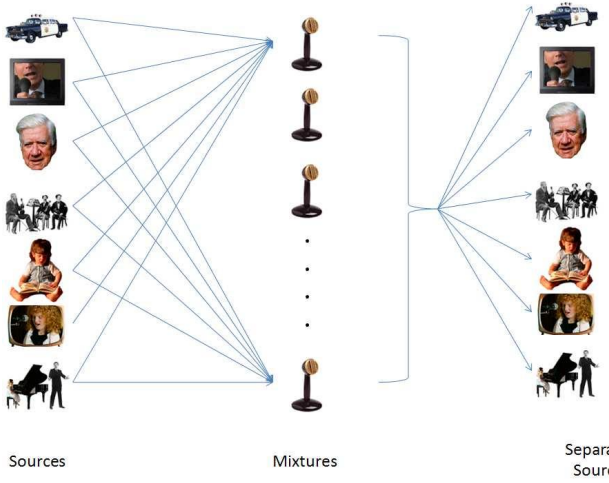


Image Compression

Fitting Machine Learning Algorithms: Cost (or Loss) Functions

“[A] cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y. This is typically expressed as a difference or distance between the predicted value and the actual value. … The objective of a ML model, therefore, is to find parameters, weights or a structure that minimises the cost function.“

- Conor McDonald

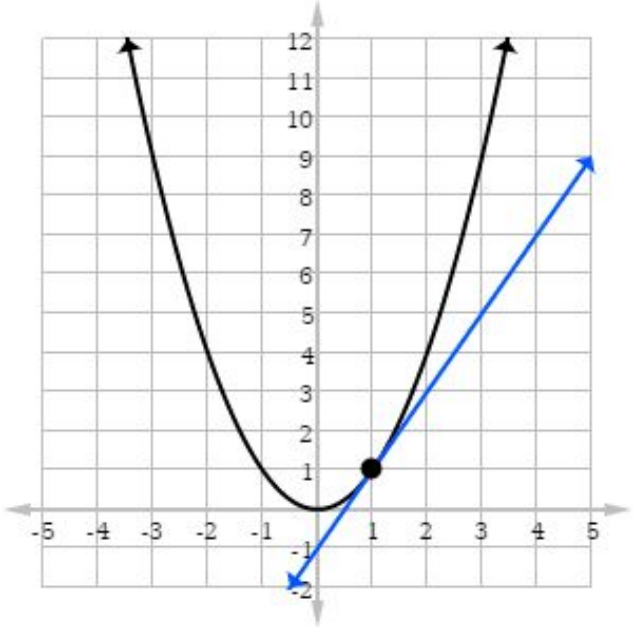
Example: Residual Sum of Squared Errors (RSS) to estimate OLS

$$RSS = \Sigma_i(y_i - \beta * X_i)^2$$

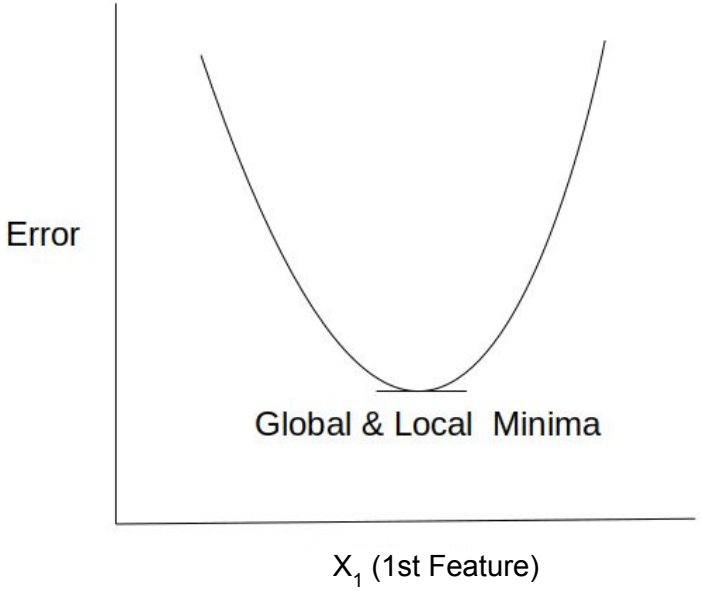
Fitting ML Algorithms: Optimization

Optimization: the selection of a best element (with regard to some criterion) from some set of available alternatives.

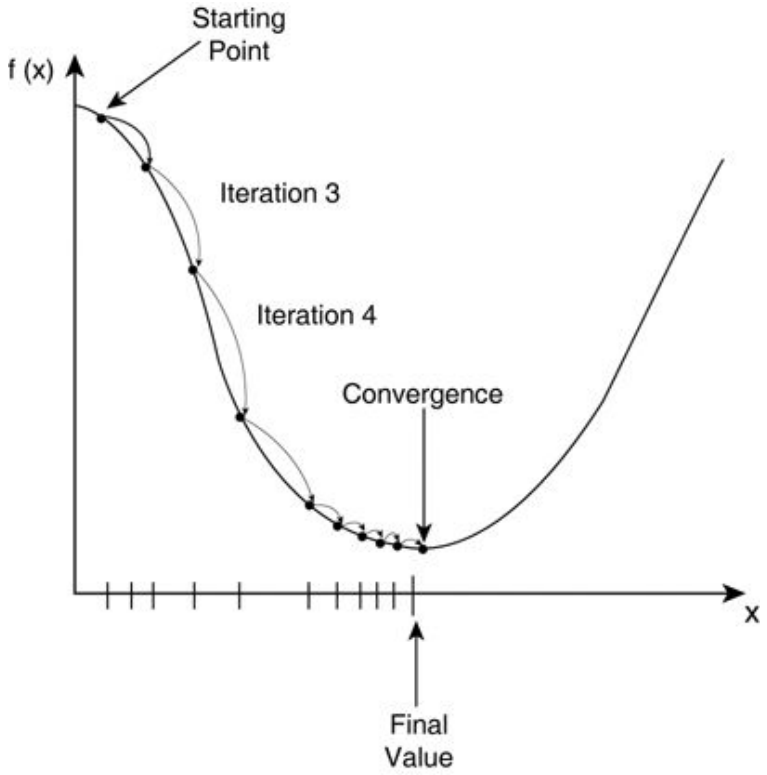
Derivative

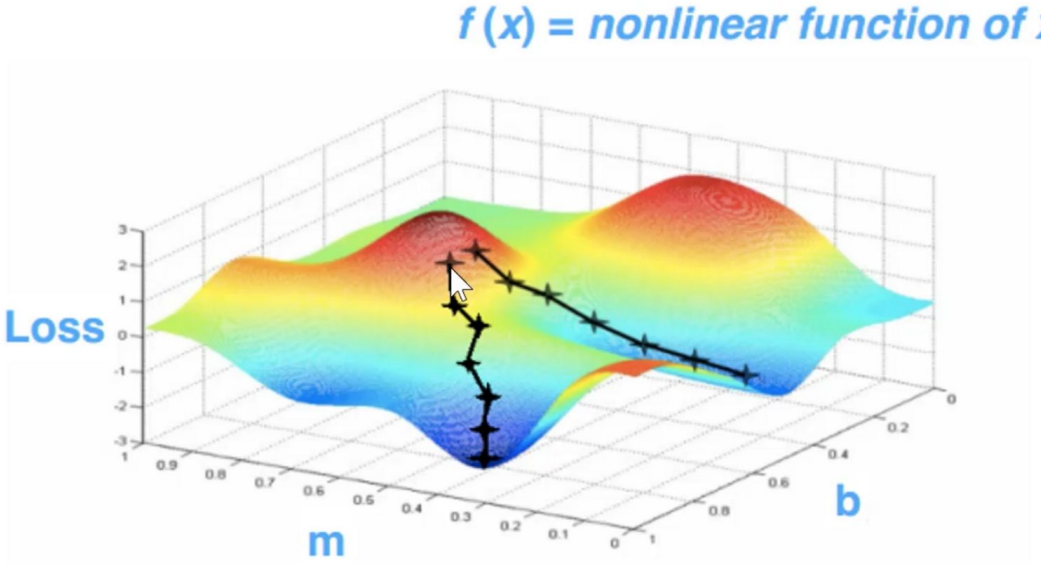


Cost Function



Fitting ML Algorithms: Gradient Descent





Model Evaluation: Classification Metrics

- **Accuracy**: The proportion of all examples correctly identified.
  - For example, the total percentage of all people that are correctly identified as either having the condition or not having it.
- **Specificity**: The proportion of actual positives that are correctly identified as such.
  - For example, the percentage of sick people who are correctly identified as having the condition.
- **Sensitivity**: The proportion of actual negatives that are correctly identified as such.
  - For example, the percentage of healthy people who are correctly identified as not having the condition.

Model Evaluation: Regression Metrics

- **Coefficient of Determination (R^2)**: The amount of variance explained in y by the model
- **Akaike Information Criterion (AIC)**: The relative amount of information lost by the model.
- **Bayesian Information Criterion (BIC)**: Similar to AIC, but penalizes each additional feature differently.

Model Evaluation: Cross-Validation

Metrics can are applied to a specific model's predictions of a dataset. That dataset can either be the training dataset used to fit the model (**in-sample testing**) or a completely different test set (**out-of-sample testing**). Cross-validation is thus a model validation technique using out-of-sample testing. It can be used to find **Hyperparameters** of models or assess model performance.

Several Types of Cross-Validation:

1. K-Fold
2. Leave-One-Out-Cross-Validation (LOOCV)
3. Held-out Test set

Model Evaluation: Overfitting and Underfitting

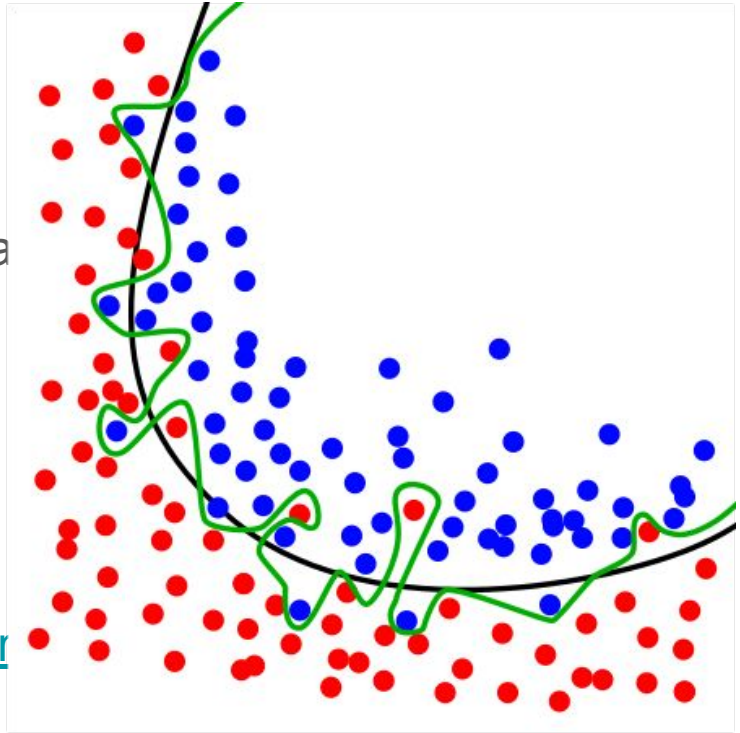
- Underfitting: When a ML “model cannot adequately capture the underlying structure of the data.”

  - Wikipedia

- Overfitting: “[T]he 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.”

  - David Leinweber

  - Cross-validation is used to reduce the effects of overfitting when validating model performance.



Further References

1. Scikit-Learn User Guide: Describes many ML models and techniques.
2. Scikit-Learn Tutorial: Full tutorials for many ML models.
3. Machine Learning and Pattern Recognition by David Bishop: Great introductory book.
4. Elements of Statistical Learning by Hastie, Tibshirani and Friedman: More advanced, but considered THE canonical book on statistical/machine learning.