Machine Learning

Algorithm design

Data

Coding

Linear Algebra

Statistics/Probability

1. Applying machine learning techniques, including:

* Predictive modeling
* Text and image mining
* Clustering
* Anomaly detection
* Forecasting methods
* Deep learning

2. Designing, development, and delivery of machine-learning-enabled solutions:

* Problem definition
* Data acquisition
* Data exploration and visualization
* Feature engineering
* Evaluating and comparing metrics

3. Working with different kinds of data:

* Structured and unstructured data sources
* Batch and streaming modes
* Formats such as tabular, image, video, audio, text, and time series

**Knowledge Required**

**A. Probability:**

1. Random variables and distributions:
   1. Expected value, variance, skewness, kurtosis
   2. Density, distribution function and quantile function
   3. Most important distributions and their properties: Gaussian, multivariate Gaussian, Student, Fisher-Snedecor, chi-square, binomial, Poisson. Gamma, Beta, exponential, Weibull might come in handy as well.
2. Central Limit Theorem. You won’t use that too often i day-to-day work, but I cannot overstress its importance in understanding what is going on.
3. Conditional probability, Bayes theorem
4. \*Conditional distributions and expected values. This is a hard one at first, but very important if you want understanding of Bayesian statistics.
5. \*Markov chains - you may never need this, but in some fields you do a lot of modeling of discrete sequential data, here Markov chains and hidden Markov models are essential.

**B. Statistics:**

1. Descriptive statistics and basic visualisations: histogram, boxplot, scatterplot, line chart, bar plot. I think these can be classified as statistical tools.
2. The basic concepts of estimation: bias, efficiency, standard errors, confidence interval. While a data scientist might need to do very little statistical testing compared to a traditional statistician role, he/she deals a lot with the concepts of uncertainty and type I and II errors, it’s very helpful to understand that.
3. Maximum likelihood and least squares estimation.
4. Linear and logistic regression (simple and multivariate)
5. Principal Components Analysis
6. Rank tests
7. Bayesian statistics, prior and posterior distribution, MCMC.
8. \*Basic time series forecasting techniques - exponential smoothing, Box-Jenkins approach. In some domains one needs to learn much more.

**C. Advanced concepts**- some of these are more helpful in research or deep understanding of theory rather than everyday tasks in commercial environments. It’s cool to know this stuff anyway:

1. Statistics as decision theory problem, loss functions
2. Statistical distances - Kullback-Leibler, Wasserstein etc. Extensively used in derivation of some deep learning methods.
3. Robust statistics - can really save your \*\*\* sometimes
4. Bootstrap
5. Kernel density estimation
6. Missing data mechanisms (MAR, MNAR, MCAR).