Probability and Linear Algebra use cases in Machine Learning

The concept of conditional independence and the invention of Bayesian networks by Judea Pearl who got the Turing award on 2011 opened a new frontier in Artificial Intelligence. Probability helped us to analyze the events both causally(cause to effect) and diagnostically(effect to cause), hereby providing intuitively correct results when rule-based system behaved incorrectly.[Pearl, J., "Reverend Bayes on inference engines: A distributed hierarchical approach," Proceedings, AAAI-82, 1982.]

Simplest Linear Model

For the simplest model $y = \theta^T x$, y is linear function of the input x parameterized by θ, probability distribution captures the uncertainty of the prediction in our stochastic world. We frame the log likelihood function and choose θ that maximizes
$\hat{\theta}_{MLE} = -(Y^{(i)} - \theta^T x^i)^2$

Spam Model

Probabilities are inherently of exponential size. Assumption of conditional independence(Naive Bayes assumption) helps to reduce the computational complexity of the probabilistic model $p_\theta(y, x_1, x_2, .., x_M)$ from O(2^m) to O(m).

Probability that a message is spam = $P(y = 1|x_1, x_2, .., x_m)$.

With Naive Bayes assumption(each random variable is conditionally independent of its non-descendent).
$P(y, x_1, x_2, .., x_m) = p(y)\prod_{i=1}^{m} p(x_i|y)$.
For spam classification we aim to estimate a function $\hat{y} = g(x)$, where y takes a discrete set of values 0,1. During prediction, we want to choose the value of y that maximizes g(x).
$g(x) = argmax_y \hat{P}(Y = y/X)$

0.1 Training

During training for Spam classification, our goal is to estimate these probabilities for all the features: $\hat{P}(Y = y/X) = \hat{P}(X/Y = y)\hat{P}(Y = y)$.
Maximum Likelihood Estimator(MLE) provides $\hat{p}(X_i = x_i/Y = y) = \frac{(No\ of\ training\ examples\ where\ X_i = x_i\ and\ Y = y)}{(No\ of\ training\ examples\ where\ Y = y)}$
LapLace Maximum A Posteriori(MAP) provides $\hat{p}(X_i = x_i/Y = y) = \frac{(No\ of\ training\ examples\ where\ X_i = x_i\ and\ Y = y)+1}{(No\ of\ training\ examples\ where\ Y = y)+2}$

0.2 Prediction

During prediction for Spam classification, we take the value of y that maximizes g(x).
$\hat{y} = g(x) = argmax_y \hat{P}(Y = y/X) = argmax_y \hat{P}(X/Y = y)\hat{P}(Y = y)$
Using Naive Bayes conditional independence, this reduces to
$= argmax_y \prod_{i=1}^{m} \hat{p}(X_i = x_i/Y = y)\hat{p}(Y = y)$
$= argmax_y \sum_{i=1}^{m} log\ \hat{p}(X_i = x_i/Y = y) + log\ \hat{p}(Y = y)$

*Disclaimer: Elaboration on Stanford University Machine Learning courses with my own understanding