Applications

With this background, let us explore how probability can apply to machine learning

Sampling – Dealing with non-deterministic processes

Probability forms the basis of sampling. In machine learning, uncertainty can arise in many ways – for example – noise in data. Probability provides a set of tools to model uncertainty. Noise could arise due to variability in the observations, as a measurement error or from other sources. Noise effects both inputs and outputs.

Apart from noise in the sample data, we should also cater for the effects of bias. Even when the observations are uniformly sampled i.e. no bias is assumed in the sampling – other limitations can introduce bias. For example, if we choose a set of participants from a specific region of the country., by definition. the sample is biased to that region. We could expand the sample scope and variance in the data by including more regions in the country. We need to balance the variance and the bias so that the sample chosen is representative of the task we are trying to model.

Typically, we are given a dataset i.e. we do not have control on the creation and sampling process of the dataset. To cater for this lack of control over sampling, we split the data into train and test sets or we use resampling techniques. ***Hence, probability (through sampling) is involved when we have incomplete coverage of the problem domain.***

Pattern recognition

Pattern recognition is a key part of machine learning. We can approach machine learning as a pattern recognition problem from a Bayesian standpoint. In Pattern Recognition – Christopher Bishop takes a Bayesian view and presents approximate inference algorithms for situations where exact answers are not feasible. For the same reasons listed above, Probability theory is a key part of pattern recognition because it helps to cater for noise / uncertainty and for the finite size of the sample and also to apply Bayesian principles to machine learning.

Training – use in Maximum likelihood estimation

Many iterative machine learning techniques like Maximum likelihood estimation (MLE) are based on probability theory. MLE is used for training in models like linear regression, logistic regression and artificial neural networks.

Developing specific algorithms

Probability forms the basis of specific algorithms like Naive Bayes classifier

Hyperparameter optimization

In machine learning models such as neural networks, hyperparameters are tuned through techniques like grid search. Bayesian optimization can be also used for hyperparameter optimization.

Model evaluation

In binary classification tasks, we predict a single probability score. Model evaluation techniques require us to summarize the performance of a model based on predicted probabilities. For example – aggregation measures like log loss require the understanding of probability theory

Applied fields of study

Probability forms the foundation of many fields such as physics, biology, and computer science where maths is applied

Probability Distribution

Probability distribution defines the likelihood of possible values that a random variable can take. PMF and PDF that have been described earlier for discrete and continuous variables respectively are probability distributions.

If we want to determine the probability distribution on two or more random variables, we use **joint probability distribution**. For a typical data attribute in machine learning, we have multiple possible values. Computing probability of all values falls under joint probability.

If we want to define the probability distribution only on a subset of variables, we use **marginal probability distribution**. This is useful if we want to estimate the probability on only a specific set of

input variables (concerning *x* attribute) when given the other input values (concerning *y* attribute).

$$P(x = X) = \sum_{y} P(x = X, y = Y)$$

For Discrete Variable

There are cases where we want to compute the probability of an event when a different event happens. This probability distribution is termed as **conditional probability distribution**.

$$P(x = X|y = Y) = \frac{P(x = X, y = Y)}{P(y = Y)}$$

Probability of x given y

Joint probability distribution can be decomposed into conditional distributions as follows:

$$P(x_1, ..., x_n) = P(x_1) \prod_{i=2}^{n} P(x_i|x_1, ..., x_{i-1})$$

Example: P(z, y, x) = P(x | y, z) * P(y | z) * P(z)

In this case, if the events are disjoint or independent, then the probability can be expressed as the product of all events' probabilities.

$$P(x = X, y = Y) = P(x = X) * P(y = Y)$$

If the events are conditionally independent, the probability is given as follows:

$$P(x = X, y = Y|z = Z) = P(x = X|z = Z) * P(y = Y|z = Z)$$

Types of Probability Distributions

Here are the distributions that we usually come across in machine learning:

1. Bernoulli Distribution

Bernoulli distribution is the probability distribution of a random variable — is *1* with a probability of *p* and 0 with a probability of *1-p*. This is typically related to a True/False or a classification scenario.

2. Binomial Distribution

Multiple Bernoulli trials constitute a binomial distribution. It's the probability distribution constituting True/False questions in *n* trials.

3. Multinoulli Distribution

Multinoulli distribution is the case where a single variable can have multiple outcomes. It's the transition from binary to several categories. When it's a multi-classification problem, this distribution comes into the picture.

4. Multinomial Distribution

Multiple Multinoulli trials constitute Multinomial distribution.

5. Gaussian Distribution

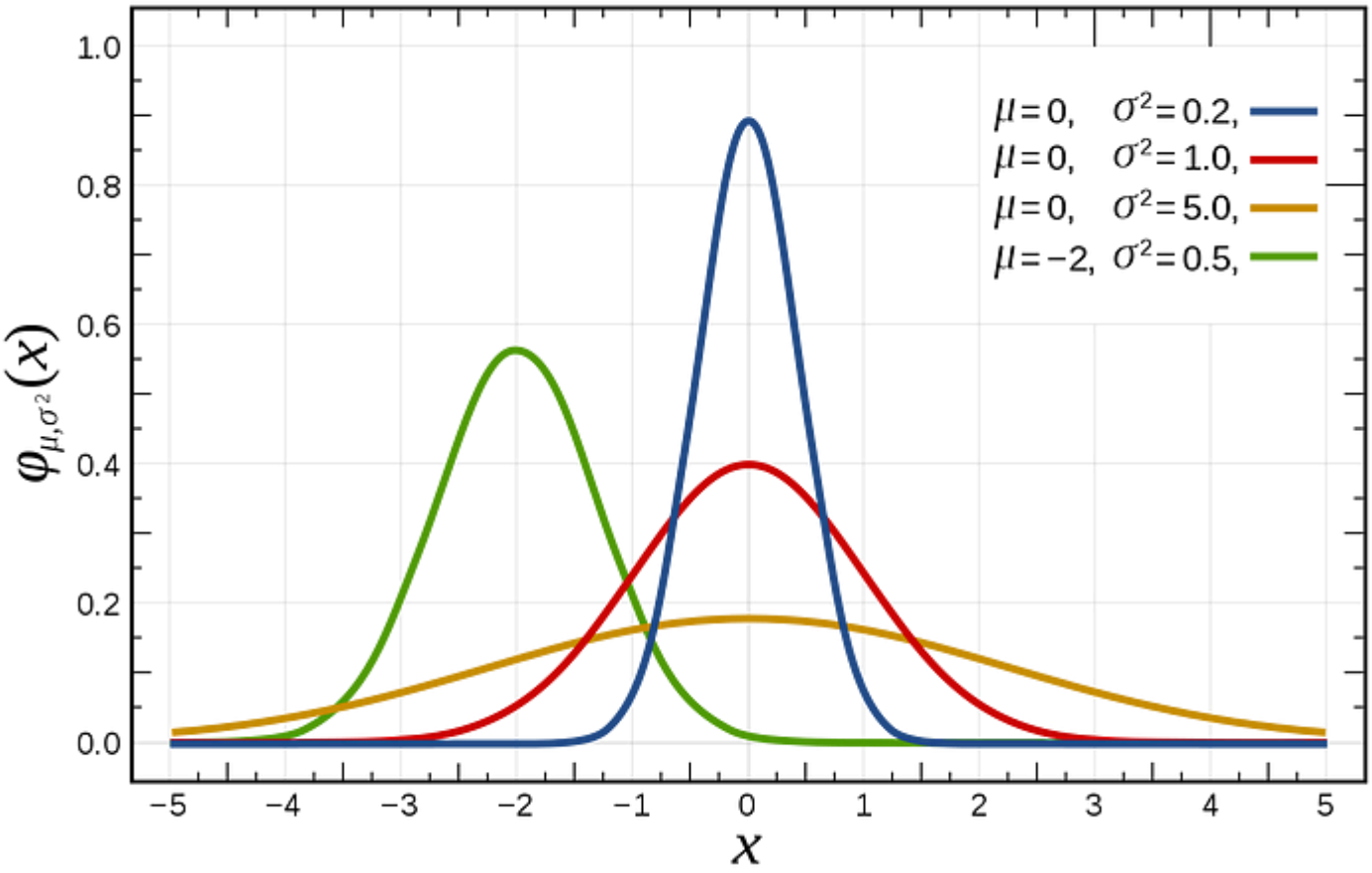
Gaussian or *Normal* Distribution is a commonly used distribution in machine learning. Several processes in our nature take the form of Gaussian distribution. In fact, there's a *Central Limit Theorem* which states that the normalized sum of several

independent variables is inclined towards Gaussian distribution irrespective of the distribution that each variable takes.

Moreover, this distribution induces maximum uncertainty into the data and requires minimum prior knowledge as it can solely be defined using the mean and variance of data.

$$N(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)}$$

mu is the mean, and sigma², the variance



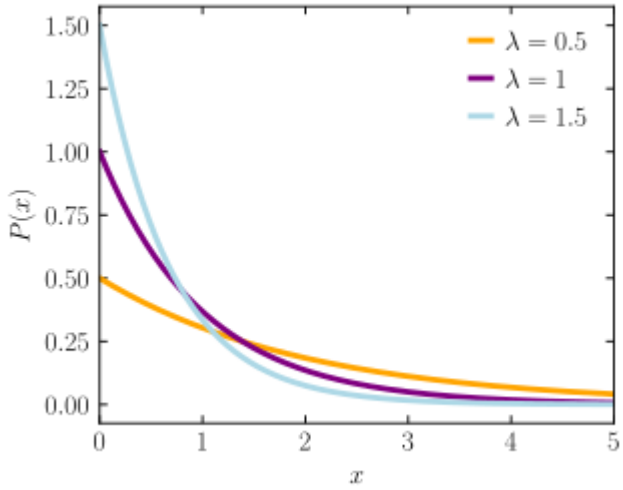
Gaussian Distribution with varying means and variances (Source: Wikipedia)

6. Exponential Distribution

Exponential distribution is concerned about the time until an event occurs. Mathematically, it has a sharp point at $x = 0$.

$$p(x;\lambda) = \lambda 1_{x\geq 0} e^{-\lambda x}$$

Probability is 0 when x < 0



Exponential Distribution at varying lambdas which is the rate parameter