Regularisation, Bayesian, MLE.

Chinthan Chandra
TA
ML T1-2023-24

Bias Variance Trade-Off

Recap:

- High Bias?
- Low Bias?
- High Variance?
- Low Variance?

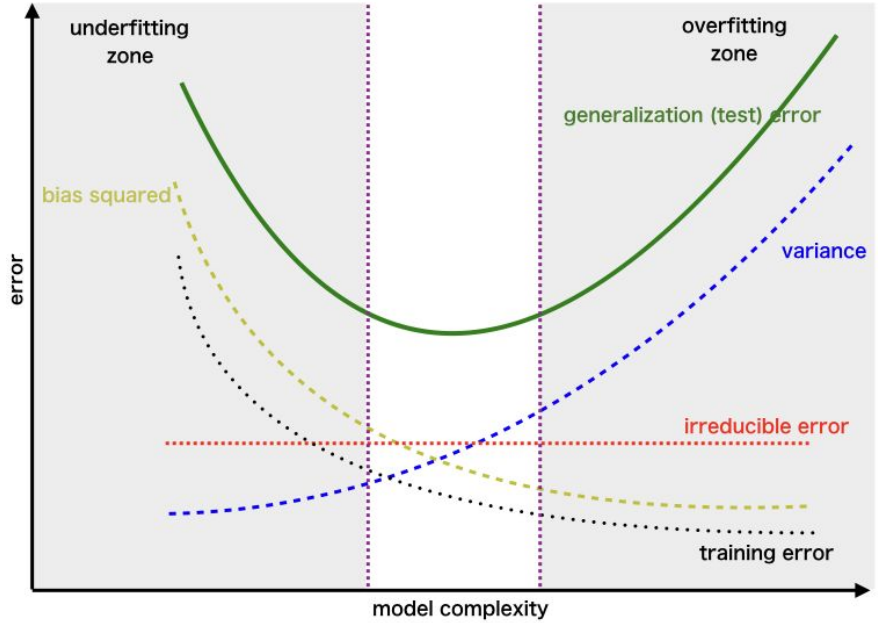
Bias Variance Trade-Off

Recap:

- High Bias- Predictions underfit data causing large prediction errors.
- Low Bias- Predictions fit data causing small prediction errors.
- High Variance- Predictions sensitive to small noise in labels.
- Low Variance- Predictions not sensitive to small noise in labels.

Optimal Model, how?

- Cross Validation.
- Training until your validation error is decreasing.
- Regularisation



Source: https://www.geeksforgeeks.org/ml-bias-variance-trade-off/

Regularization

Techniques that ensure the model does not overfit to the training data.

How? A penalty is added to the model's loss function.

Two main types:

1. Lasso (L1): $\text{LossModified} = \text{Loss} + \lambda \|w\|$
2. Ridge (L2): $\text{LossModified} = \text{Loss} + \lambda \|w\|^2$

Many others exist, few of them on scikit-learn.

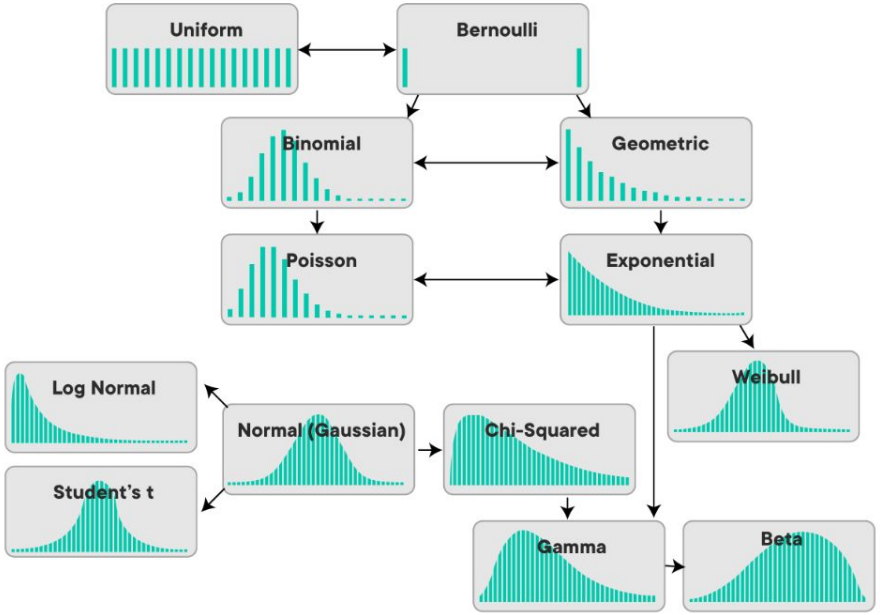
Distributions

- How is data distributed?

Various types of distribution exist. Gaussian(Normal), Bernoulli, Binomial, would of some of them which you have heard.

- Why distributions?

Most of our data will have some form of distribution and our aim here is to find a function which describes the distribution.



Source: Link

Normal Distribution

What is a normal distribution?

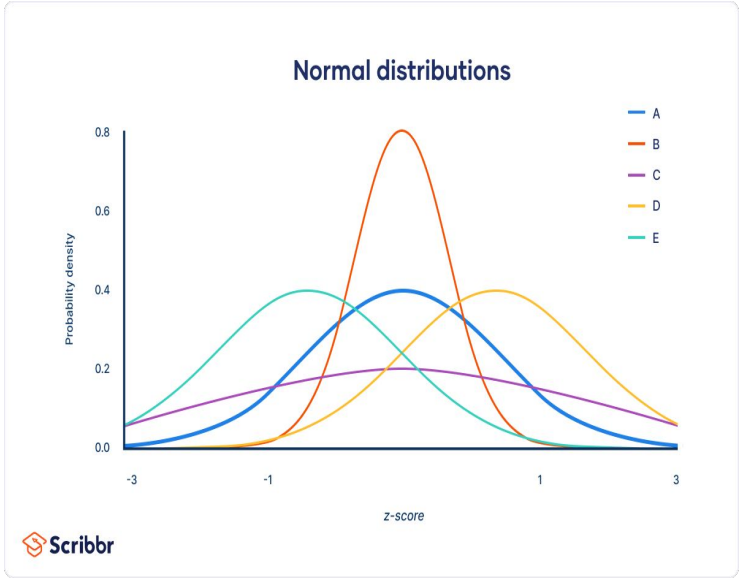
The best example would be the marks distribution of class. Where most of the students have average marks and the other two ends have lesser density.

Given by the formula,

$$N(\sigma, \mu) \sim \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$

Why a Normal Distribution?

Most of the real world data can be converted to a normal distribution. It makes math easier.



Source: Link

Likelihood

What is likelihood?

Given two random variables X and Y.

If we have given value for one of them, the probability that the other variable will take a certain value is called as likelihood.

$$L(\theta \mid x) = P(X = x \mid Y = \theta) = \frac{P(Y=\theta \mid X=x) \cdot P(X=x)}{P(Y=\theta)}$$

In Machine Learning we use this to see how likely is to see an event occur given the data point i.e.

$$P(Y = \theta \mid X = x) = \frac{P(X = x \mid Y = \theta) \cdot P(Y = \theta)}{P(X = x)}$$

Maximum Likelihood Estimation

Given a dataset, we have many data points, our aim is to maximise the likelihood for every data point.

Why?

How do we do it? $$L(\theta \mid x) = P(X = x \mid Y = \theta) = \frac{P(Y=\theta \mid X=x) \cdot P(X=x)}{P(Y=\theta)}$$

In the above equation, we have defined it for one data point. For all data points, the likelihood would be the product of them.

$$L(\theta \mid X) = \prod_{c_j} \prod_{\theta_j = c_j} P\Big(X = x_i \mid Y = \theta_j\Big)$$

So we take the product over all data points for all the values they can take.

Maximum Likelihood Estimation

$$L(\theta \mid X) = \prod_{c_j} \prod_{\theta_j = c_j} P\Big(X = x_i \mid Y = \theta_j\Big)$$

If our distribution was a Bayesian Distribution.

We know that, $P(X = x_i \mid Y = \theta_j) = N\big(\sigma_{\theta_j}, \mu_{\theta_j}\big)$

So our likelihood function changes to products of Normal Distributions.

How do we maximise this? Take log first. $\log L(\theta \mid X) = \sum_{c_j} \sum_{\theta_j = c_j} \left(- \log\left(\sigma_{\theta_j}\right) - \frac{1}{2} \cdot \frac{\left(x_i - \mu_{\theta_j}\right)^2}{\sigma_{\theta_j}^2} \right)$

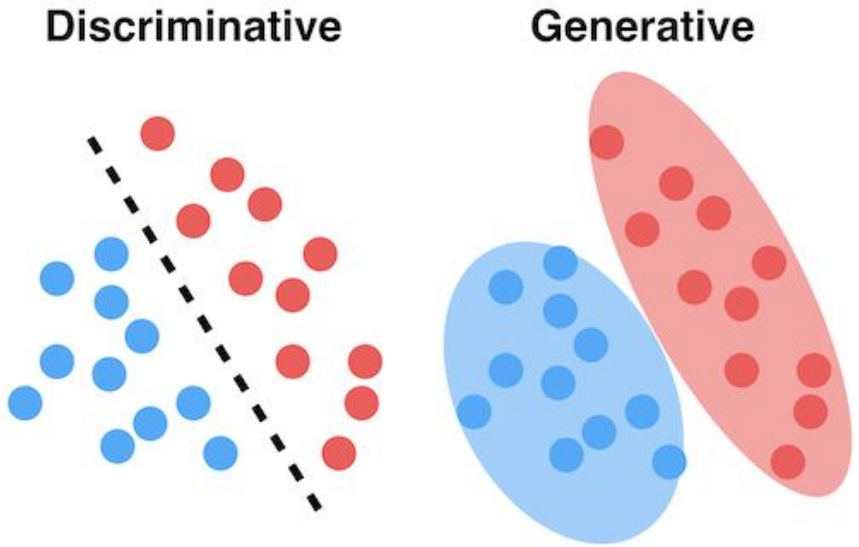
This is called the log likelihood.

Now we need to differentiate with the mean and the variance.

Generative vs Discriminative

Generative models are those models which are able to produce new data points similar to the existing data points in the training data.

Discriminative models usually have discriminating boundaries. They can not generate new data points.



Source: Link

Thank You