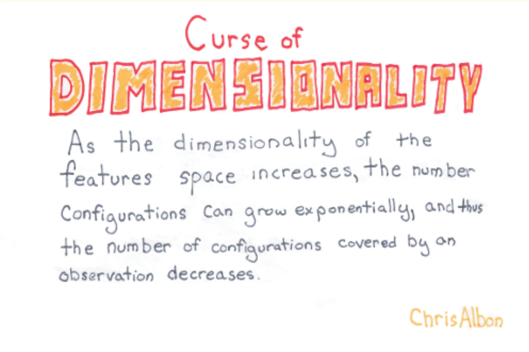
Curse of Dimensionality:

→ ML excels at analyzing data with many dimensions, but it becomes challenging to create meaningful models as the number of dimensions increases.

→ curse of dimensionality

↳ increasing data dimensions and its explosive tendencies

→ increase computational efforts for analyze and process the data.

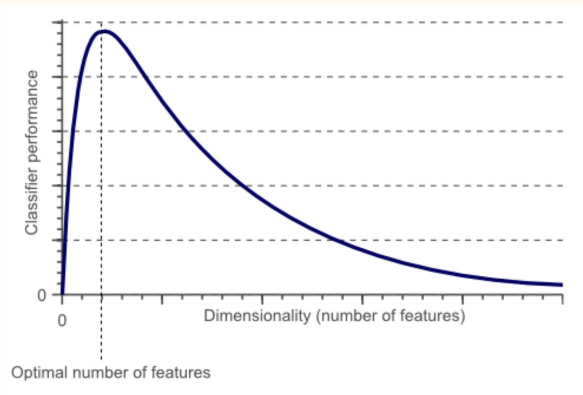


→ say, # of features = p
# of data points = n

↳ CoD ⇒ p >> n

Hughes Phenomenon:

→ say that model's performance increases with the increasing number of features until we reach the optimal # of features.



→ distance = Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

→ each new dimension adds a non-negative term to the sum.

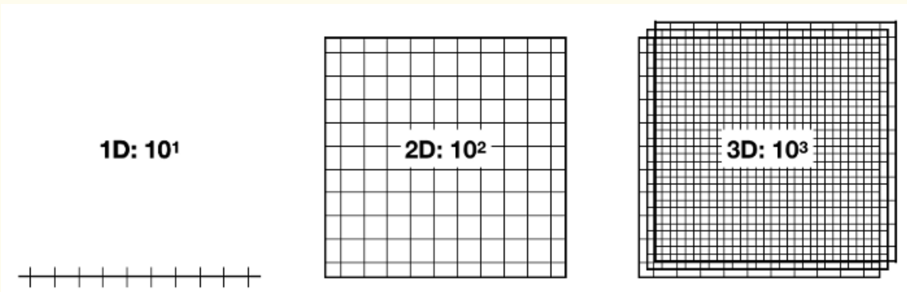
↳ distance ↑ $\propto$ # dimension in feature vector ↑

→ now with given datapoints -

as # of dimension ↑ ⇒ feature space becomes extremely sparse → less dense or emptier

→ In other words,

↳ lower data density requires more observations to keep avg distance b/w datapoints same.

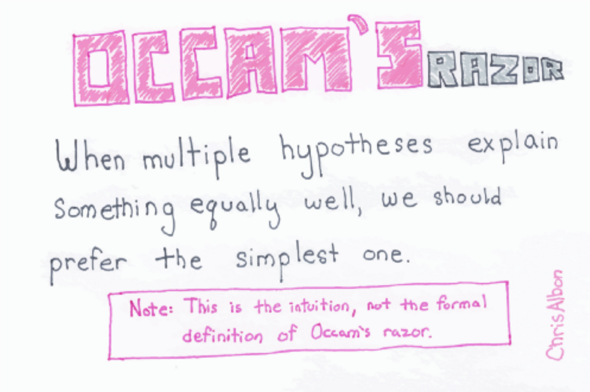


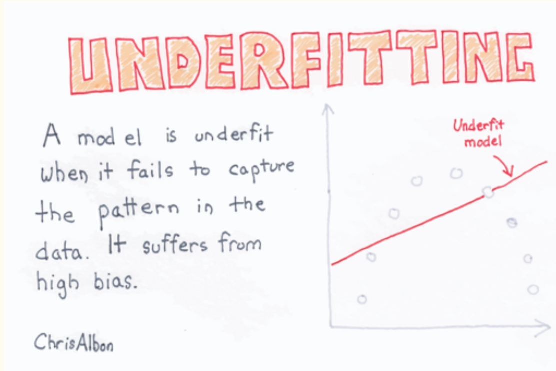
The number of features required to keep average distance constant grows exponentially with the number of dimensions.

Overfitting:

- variance increases as they get more opportunity to overfit to noise in more dimensions.
  - ↳ poor generalization
- KNN is very susceptible to overfitting due to CoD.
  - ↳ sparsity ↑ with dimension ↑ for fixed # of datapoints.
  - ↳ so the closest neighbour is also too far away in high dim space to give a good estimate.
- There is no scope of adding regularization in KNN so it suffers from CoD.
- as $p >> n$ ⇒ datapoints << features
  - ↳ overfit
    - ↳ fails in test data.

Underfit:





→ to reduce chance of overfitting, we can prefer simple model with less parameters.

↳ But this may suffer from underfitting

solution:

→ Forward feature selection

→ Dimension Reduction

↳ PCA

↳ t-SNE