Train In Data

Mutual Information
Continuous variables

Mutual Information - discrete

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)} \right)$$

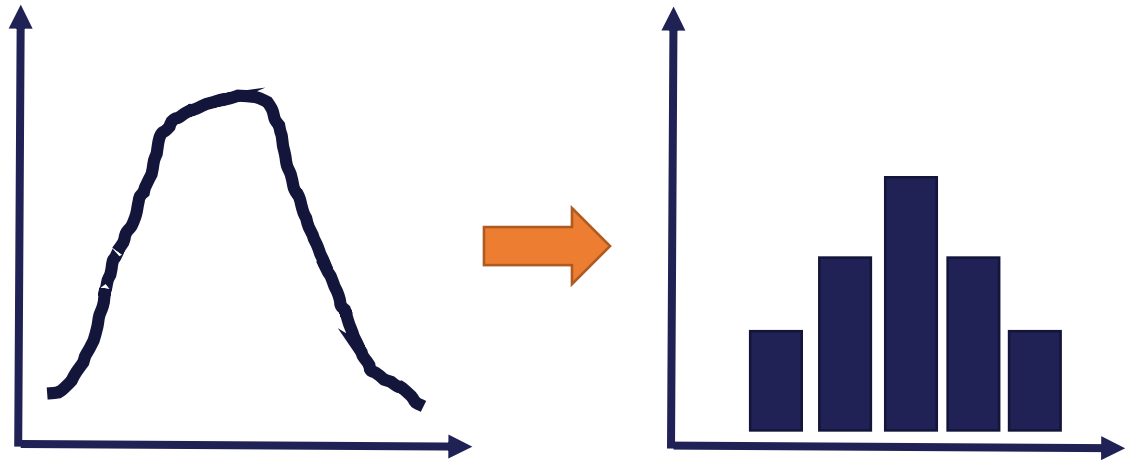
- I is the mutual information
- p(X,Y) is the joint probability of X and Y
- p(X) is the probability of X
- p(Y) is the probability of Y

Mutual Information - continuous

$$I(X,Y) = \int_X \int_Y p(x,y) log(\frac{p(x,y)}{p(x)p(y)})$$

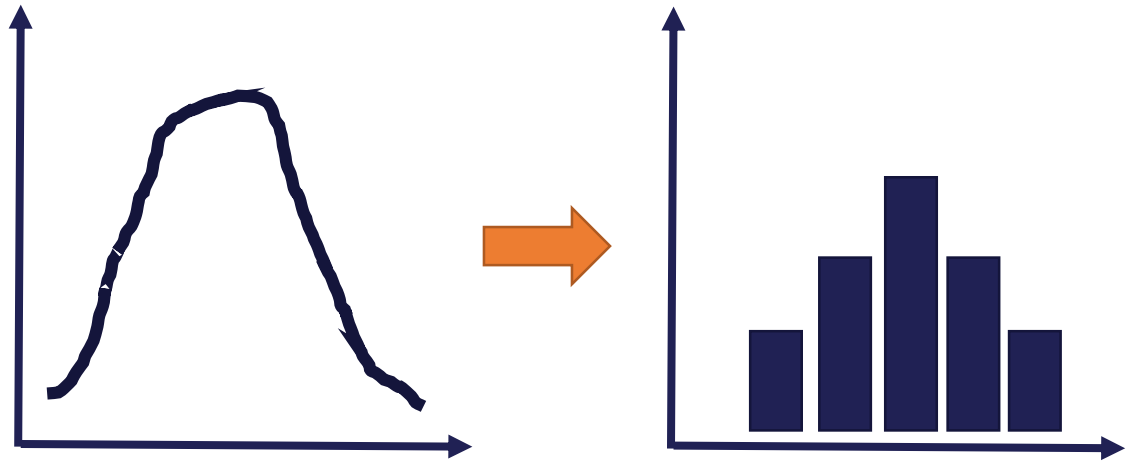
- Replace summation with integrals
- We don't know p(X) and p(Y), so we have to infer them
- How?

Mutual Information - continuous



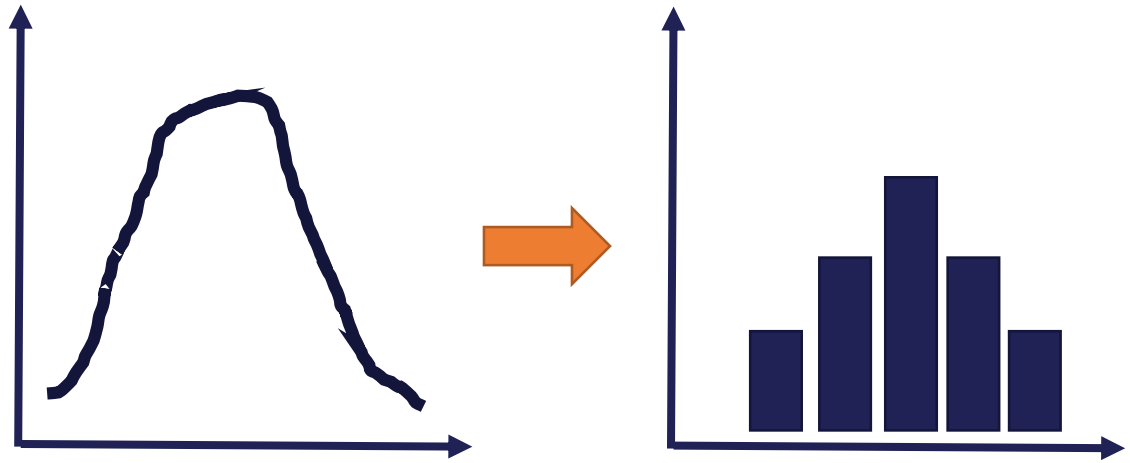
- Discretize the continuous variable and proceed as normal.

Mutual Information - continuous



- Discretize the continuous variable and proceed as normal.
- **Equal width or equal frequency bins?**
- **How many?**

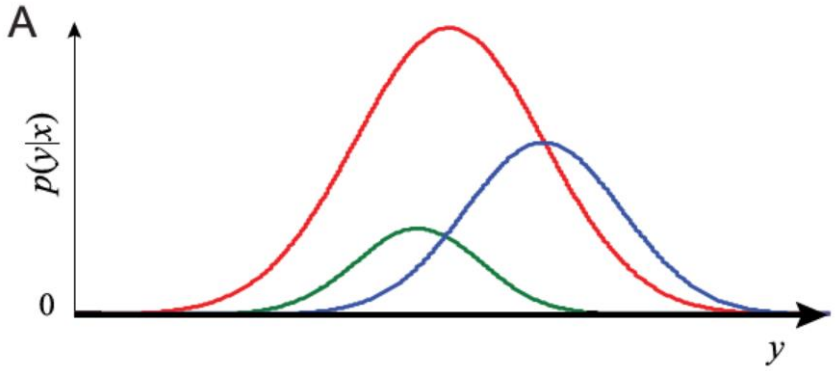
Mutual Information - continuous



- Discretize the continuous variable and proceed as normal.
- **Equal width or equal frequency bins?**
- **How many?**

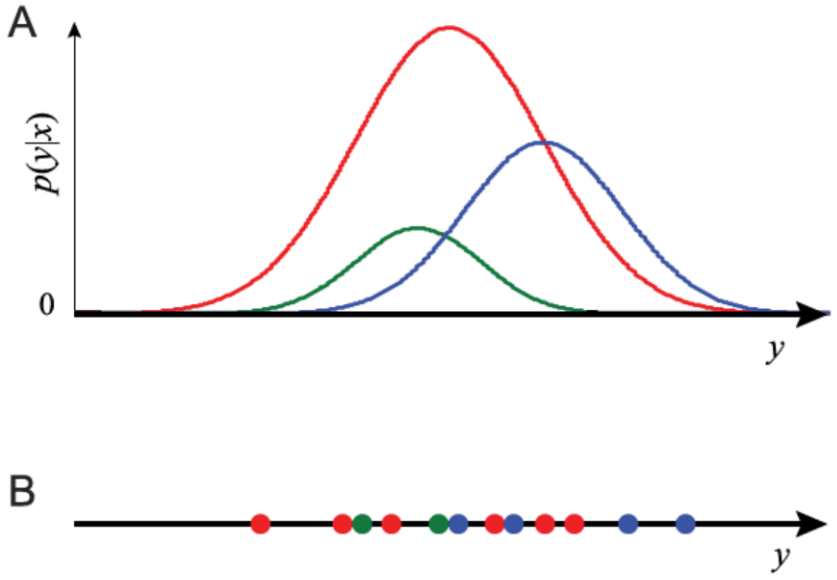
- Not what sklearn does.
- Sklearn uses the KNN approach

Use of KNN to estimate MI



- Y ➔ continuous variable (x axis)
- X ➔ discrete variables (colours)

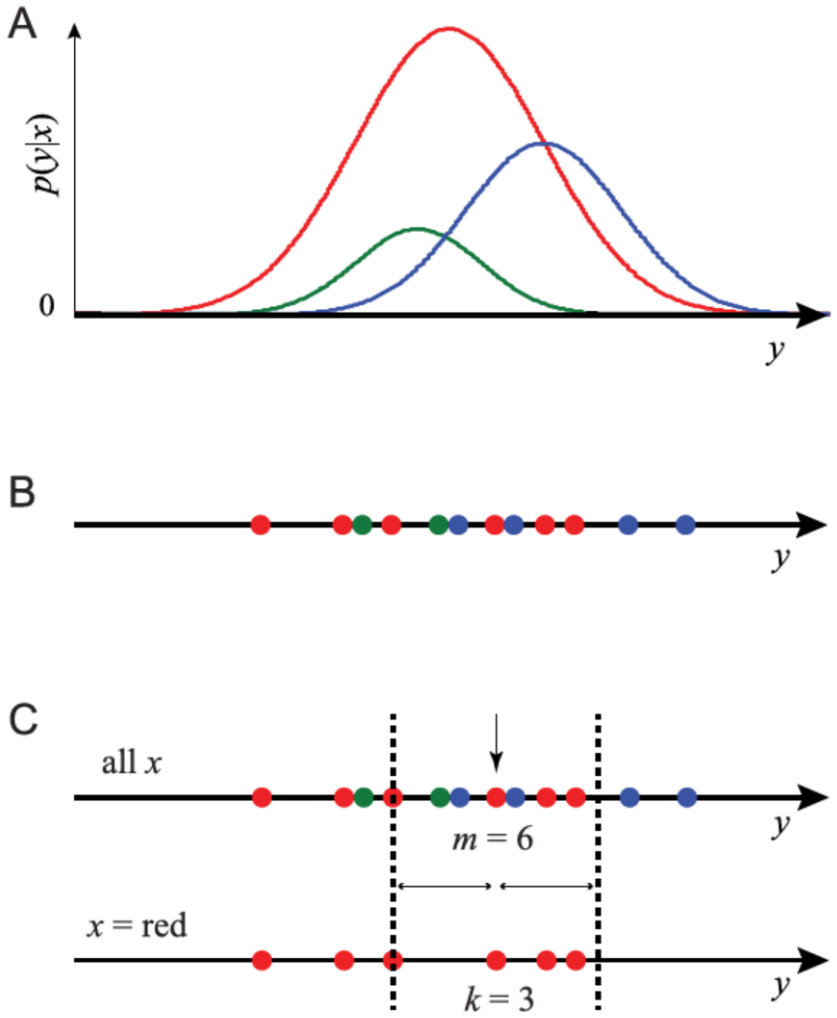
Use of KNN to estimate MI



- Y ➔ continuous variable (x axis)
- X ➔ discrete variables (colours)

- Sample at random (**N=12**)
- We'll start with red (**N_i = 6**)

Use of KNN to estimate MI



- Y ➔ continuous variable (x axis)
- X ➔ discrete variables (colours)

- Sample at random (**N=12**)
- We'll start with red (**N_i = 6**)

- Find the k closest friends of one sample (**k=3**)
- Find the distance to the furtherst neighbour
- Count all samples within that interval (**mi=6**)

Use of KNN to estimate MI

$$I_i = \psi(N) - \psi(N_{xi}) + \psi(k) - \psi(m_i)$$

ψ is the digamma function

To estimate the MI from the data set, we average I_i over all data points.

- Y ➔ continuous variable (x axis)
- X ➔ discrete variables (colours)

- Sample at random (**N=12**)
- We'll start with red (**N_i = 6**)

- Find the k closest friends of one sample (**k=3**)
- Find the distance to the furtherst neighbour
- Count all samples within that interval (**mi=6**)

Use of KNN to estimate MI

- KNN method has been shown to give more accurate values of MI
- For 2 continuous variables, the closest neighbours are sampled within a “radius” of the observation being examined.

Train In Data

THANK YOU

www.trainindata.com