Classification and the Statistical Learning Framework:

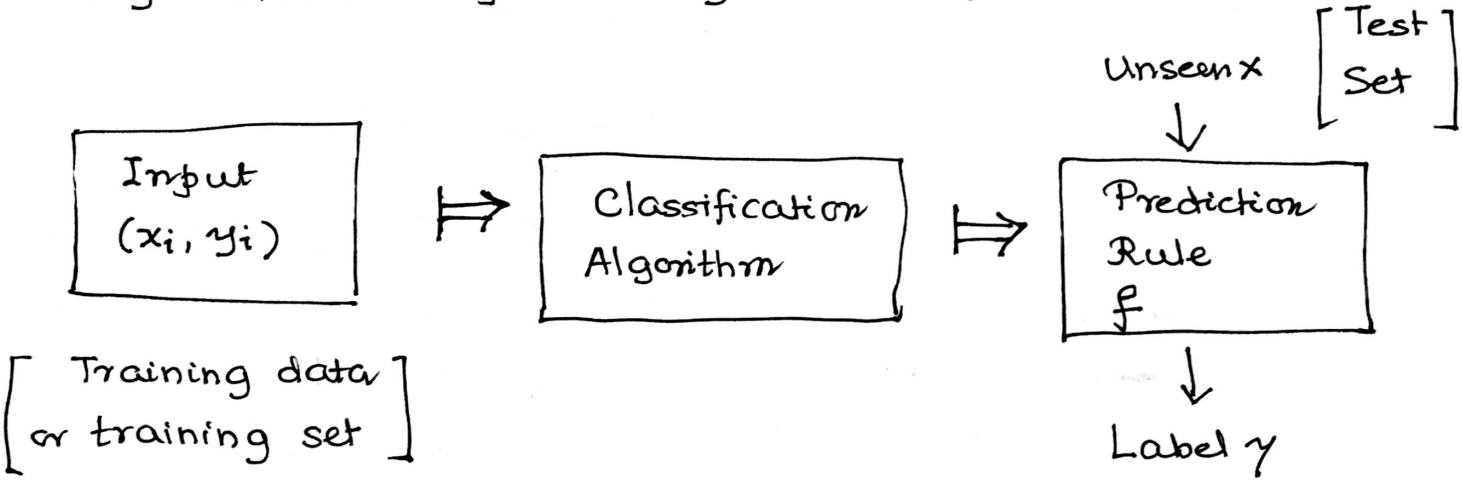
Given labelled data

(x_i, y_i) $\quad i = 1, .., n$

x_i: feature vectors

y_i: labels (discrete)

design a rule to predict y values for unseen x.



Performance Measures:

1. Training error: If f is the prediction rule,

$$= \frac{\#\text{times } f(x_i) \neq y_i \text{ on the training set}}{\text{size of training set}}$$

2. Test error := $\dfrac{\#\text{times } f(x_i) \neq y_i \text{ where } (x_i, y_i) \text{ is in test set}}{\text{size of test set}}$

- Training and test data MUST be kept separate
- Test error is a better measure than training error
- Test and training data should be "similar"

The Statistical Learning Framework:

Assumption: All data (training, test, etc) is drawn iid from some unknown underlying distribution D.

D: called the **data distribution**

$\mathcal{X}$: space of feature vectors

$\mathcal{Y}$: set of all labels

D is a distribution over $\mathcal{X} \times \mathcal{Y}$

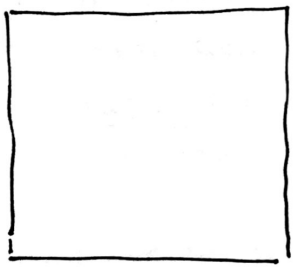
How to sample from D?

1. Draw $(x, y) \sim D$
2. Draw y according to its marginal distribution, then draw x according to the conditional distribution of $x|y$
3. Draw x according to its marginal distribution, then y from the conditional distribution of $y|x$.

μ: distribution on $\mathcal{X}$

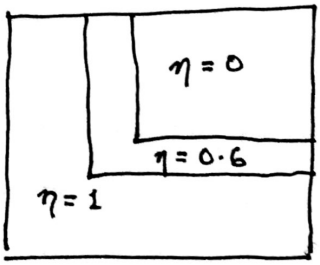
η: conditional distribution of $y|x$.

Eg:



μ: uniform on square

$\eta(x) := P(y = 1 | x)$



Why isn't $\eta(x) = 0$ or 1?

- sometimes it is, but sometimes there is inherent uncertainty
- this happens when the features are not enough to predict the label.

Example: Age, Presence of Gene A (features) → Disease or Not (Label)

Just because someone has a gene doesn't mean they have a disease!

Limitations of Statistical Learning Framework:

Sometimes assumption does not hold.

* μ can change.
* μ and η can both change.

Examples:

* Training data is data on whether offenders given bail have reoffended or not. Tested on new offenders.
  Here training data distribution is different from test
  - ↓ conditioned on given bail. Maybe minorities were not given bail
  - ↓ entire population of offenders [Raises ethical concerns]
* Task is to predict topic of news, based on words in it. With time both distribution of x changes, and also distribution of $y|x$ → Donald Trump → business / politics