Classification and the Statistical Learning Framework:

Given labelled data

\[(x_1, y_1), \ldots, (x_n, y_n)\]

\[\downarrow \quad \downarrow\]

feature vectors
labels (discrete)

design a rule to predict y value for unseen x.

Performance Measures:

1. Training error: If f is the prediction rule,
   \[
   \mathbb{P}(f(x_i) \neq y_i \text{ on the training set}) / \text{size of training set}
   \]

2. Test error: \[
   \mathbb{P}(f(x_i) \neq y_i \text{ where } (x_i, y_i) \text{ is in test set}) / \text{size of test set}
   \]

- 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
- \( X \): space of feature vectors
- \( Y \): set of all labels
- \( D \) is a distribution over \( X \times Y \)

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 \( x \mid y \)
3. Draw \( x \) according to its marginal distribution, then \( y \) from the conditional distribution of \( y \mid x \).

\( \mu \): distribution on \( x \)
\( \eta \): conditional distribution of \( y \mid x \).

Eq.:

\[
\eta(x) := P(y = 1 \mid x)
\]

\( \mu \): uniform on square
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 → Disease or Not

Features

Label

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

Limitations of Statistical Learning Framework:
Sometimes assumption does not hold.

* \( \mu \) can change.
* \( \mu \) and \( \eta \) 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 condition on given bail
  Maybe minorities were not given bail

\[ \downarrow \]

entire population
of offenders
[Raises ethical concerns]

* Task is to predict topic of news, based on Nords in it.
  With time both distribution of \( x \) changes, and also distribution of \( y | x \) → Donald Trump → business
  → politics