Inductive Learning

What we know about learning so far

- Using unseen test data to measure/compare the performance of the learning algorithms
- The choice of the performance measure depends on the problem to be solved
- The training data and the test data should be related (i.e., I cannot tell you machine learning and ask you about literature!).

Let's formalize the idea of performance measure

- We want the performance measure to tell us how different/bad a system's prediction is in comparison to the truth
- In machine learning, we achieve this via the loss functions to quantify the difference $L(\hat{y}, y)$
  - Regression:
    - Squared loss: $L(\hat{y}, y) = (\hat{y} - y)^2$
    - Absolute loss: $L(\hat{y}, y) = |\hat{y} - y|$
  - Binary classification:
    - Zero/one loss: $L(\hat{y}, y) = \begin{cases} 0 \; if \; \hat{y} = y \\ 1 \; otherwise \end{cases}$

Let's formalize the relatedness of training and test data

- Data can be seen as being drawn/sampled from a probability distribution.
- In machine learning, we use the probabilistic model of learning: the training data and test data should come from the same distribution $\mathcal{D}$ (i.e., the data generating distribution)
- The distribution $\mathcal{D}$ is defined over the pairs of input and output (x, y):

$$\mathcal{D} = P(x, y)$$

$$(x, y) \sim P(x, y)$$

The data generating distribution

- No assumption is made on $\mathcal{D}$
- $\mathcal{D}$ is typically unknown (the fundamental problem in machine learning)
- We often only have a sample (called training data):

$D = \{(x_1, y_1), (x_2, x_2), \ldots, (x_n, y_n)\}$ from $\mathcal{D}$

leading to an empirical distribution over D:

$$\hat{P}(x, y) = \frac{number\ of\ times\ (x,y)\ appears\ in\ D}{|D|}$$

The learning problem

- Ideally, we want to find a mapping function f that has minimal error on every possible input-output pairs ($f: x \longrightarrow y$)
- Formally, using the loss function and the data generating distribution, this translates into minimizing:

$$f^* = argmin_f \epsilon(f)$$

$$\epsilon(f) = E_{(x,y)\sim P(x,y)}[L(f(x), y)]$$

The learning problem

- However, as we only have the training data D in practice, we need to approximate:

$$P(x, y) \approx \hat{P}(x, y)$$

$$\epsilon(f) = E_{(x,y)\sim P(x,y)}[L(f(x), y)]$$

$$\approx E_{(x,y)\sim \hat{P}(x,y)}[L(f(x), y)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i))$$

Inductive bias

- However, many different functions f might have the same training error.
- Different models would specify their own preference over such functions, called inductive bias

Without bias we cannot learn!

What are the biases in ID3 for decision trees?

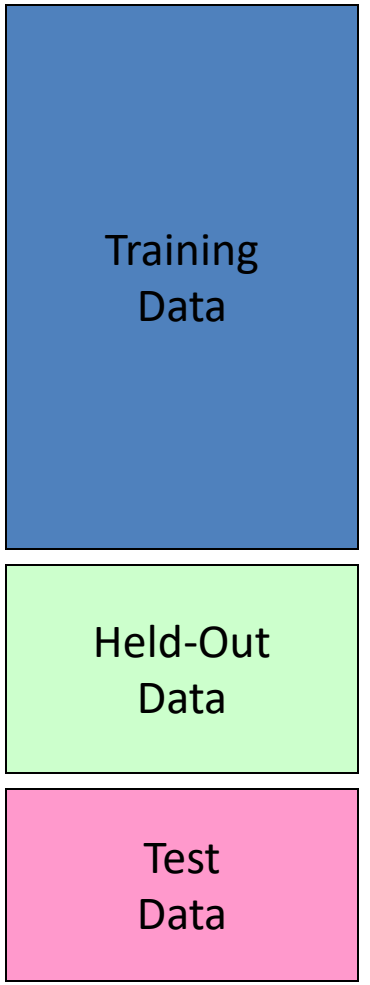
Which biases human has when learning?

Tuning and Testing

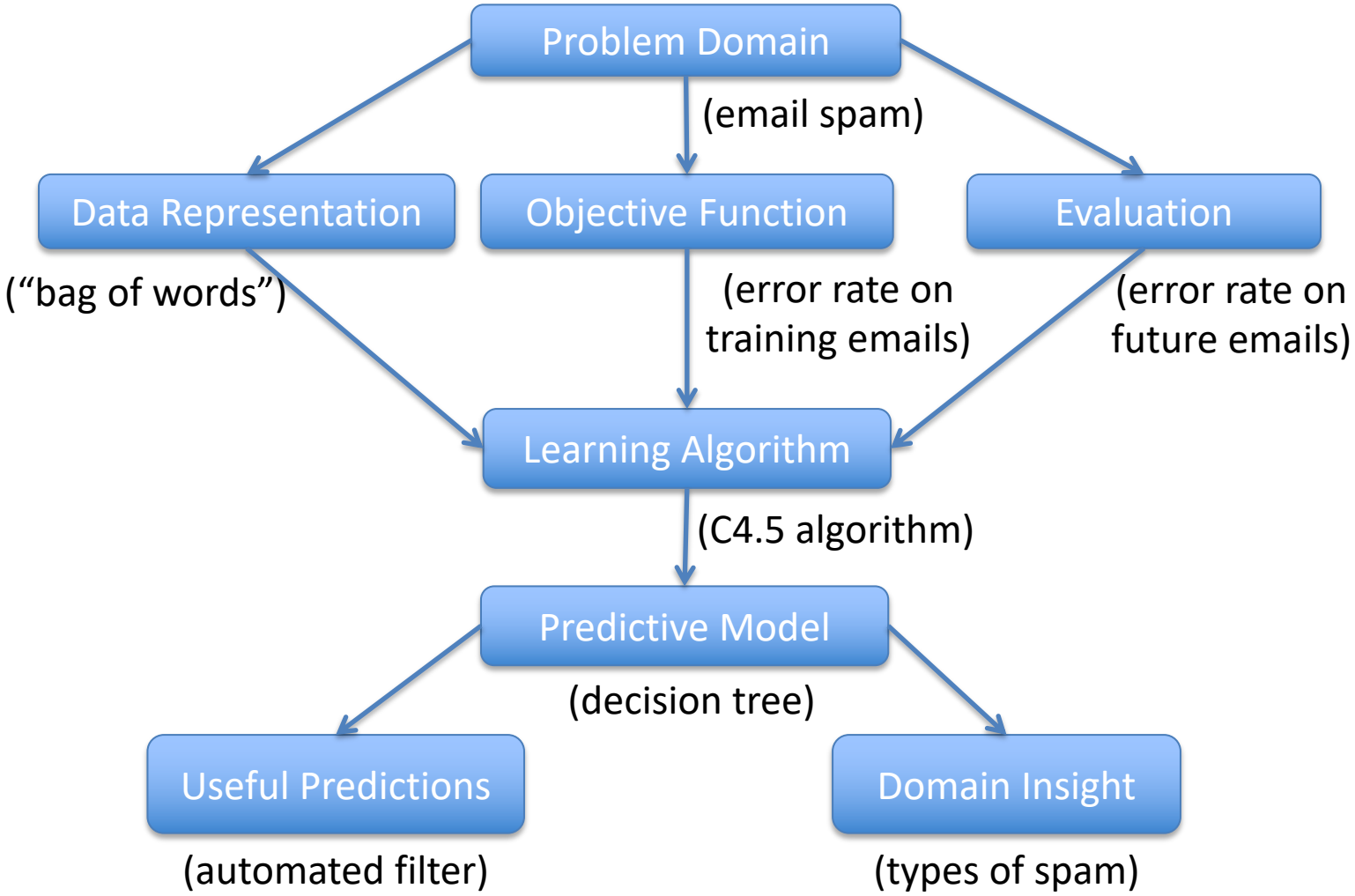
- How do we know if *f* is good?
  - Can't use the data generating distribution directly
  - Low training error could be misleading – easy to memorize data and overestimate performance.
- Average loss on previously unseen data is a much better indicator of future performance.
  - Use held out validation data to choose algorithms and hyperparameters (algorithm settings) that are likely to generalize well.
  - Use separate test data for final evaluation.

Important Concepts

- Data: labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set
  - Test set
- Features: attribute-value pairs which characterize each x
- Experimentation cycle
  - Learn parameters (e.g. model probabilities) on training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never "peek" at the test set!
- Evaluation
  - Accuracy: fraction of instances predicted correctly
- Overfitting and generalization
  - Want a classifier which does well on *test* data
  - Overfitting: fitting the training data very closely, but not generalizing well



The Big Picture



Not everything is learnable

- Noise in training data
- The provided features are insufficient for learning
- One input might have more than one correct output (e.g., offensive vs non-offensive)
- Inductive bias is misaligned with the concept being learned (can be fixed by choosing a different learning method).