The Bayes Optimal Classifier

Machine Learning
Fall 2017

Most probable classification

- In Bayesian learning, the primary question is: What is the most probable hypothesis given data?
- We can also ask: For a new test point, what is the most probable label, given training data?
- Is this the same as the prediction of the maximum a posteriori hypothesis?

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis?

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? h_1

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? h_1
- For a new instance **x**, suppose $h_1(\mathbf{x}) = +1$, $h_2(\mathbf{x}) = -1$ and $h_3(\mathbf{x}) = -1$

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? h_1
- For a new instance **x**, suppose $h_1(\mathbf{x}) = +1$, $h_2(\mathbf{x}) = -1$ and $h_3(\mathbf{x}) = -1$
- What is the most probable classification of **x**?

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? h_1
- For a new instance **x**, suppose $h_1(\mathbf{x}) = +1$, $h_2(\mathbf{x}) = -1$ and $h_3(\mathbf{x}) = -1$
- What is the most probable classification of **x**? -1

$P(+1 \mid \mathbf{x}) = 0.4$ $\quad$ $P(-1 \mid \mathbf{x}) = 0.3 + 0.3$

Most probable classification

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? h_1
- For a new instance **x**, suppose $h_1(\mathbf{x}) = +1$, $h_2(\mathbf{x}) = -1$ and $h_3(\mathbf{x}) = -1$
- What is the most probable classification of **x**? -1

$P(+1 \mid \mathbf{x}) = 0.4$ $\quad$ $P(-1 \mid \mathbf{x}) = 0.3 + 0.3$

The most probable classification is not the same as the prediction of the MAP hypothesis

Bayes Optimal Classifier

- How should we use the general formalism?
  - What should H be?

Bayes Optimal Classifier

- How should we use the general formalism?
  - What should H be?

H can be a collection of functions.

- Given the training data, choose an optimal function
- Then, given new data, evaluate the selected function on it

H can be a collection of possible predictions

- Given the data, try to directly choose the optimal prediction

Bayes Optimal Classifier

- How should we use the general formalism?
  - What should H be?

H can be a collection of functions.

- Given the training data, choose an optimal function
- Then, given new data, evaluate the selected function on it

H can be a collection of possible predictions

- Given the data, try to directly choose the optimal prediction

These two could be different!
Selecting a function vs. entertaining all options until the last minute

Bayes Optimal Classification

Defined as the label produced by the most probable classifier

$$\arg\max_{y} \sum_{h_i \in H} P(y|h_i)P(h_i|D)$$

Computing this can be hopelessly inefficient

And yet an interesting theoretical concept because, no other classification method can outperform this method on average

(using the same hypothesis space and prior knowledge)