Supervised Learning

Given:

- a set of inputs features $X_1, \ldots, X_n$
- a set of target features $Y_1, \ldots, Y_k$
- a set of training examples where the values for the input features and the target features are given for each example
- a new example, where only the values for the input features are given

predict the values for the target features for the new example.

Supervised Learning

Given:

- a set of inputs features $X_1, \ldots, X_n$
- a set of target features $Y_1, \ldots, Y_k$
- a set of training examples where the values for the input features and the target features are given for each example
- a new example, where only the values for the input features are given

predict the values for the target features for the new example.

- classification when the Y_i are discrete
- regression when the Y_i are continuous

Example Data Representations

A travel agent wants to predict the preferred length of a trip, which can be from 1 to 6 days. (No input features).

Two representations of the same data:

— Y is the length of trip chosen.

— Each Y_i is an indicator variable that has value 1 if the chosen length is i, and is 0 otherwise.

Example	Y
e_1	1
e_2	6
e_3	6
e_4	2
e_5	1

Example	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
e_1	1	0	0	0	0	0
e_2	0	0	0	0	0	1
e_3	0	0	0	0	0	1
e_4	0	1	0	0	0	0
e_5	1	0	0	0	0	0

What is a prediction?

Evaluating Predictions

Suppose we want to make a prediction of a value for a target feature on example e:

- o_e is the observed value of target feature on example e.
- p_e is the predicted value of target feature on example e.
- The error of the prediction is a measure of how close p_e is to o_e.
- There are many possible errors that could be measured.

Sometimes p_e can be a real number even though o_e can only have a few values.

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$
- number wrong: $L_0(E) = \#\{e : o_e \neq p_e\}$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$
- number wrong: $L_0(E) = \#\{e : o_e \neq p_e\}$
- A cost-based error takes into account costs of errors.

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e} (1 - p_e)^{(1 - o_e)}$$

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e} (1 - p_e)^{(1 - o_e)}$$

- log likelihood

$$\sum_{e \in E} (o_e \log p_e + (1 - o_e) \log(1 - p_e))$$

  is negative of number of bits to encode the data given a code based on p_e.