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

Evaluating Predictions

Suppose F is a feature and e is an example:

- val(e,F) is the value of feature F for example e.
- pval(e,F) is the predicted value of feature F for example e.
- The error of the prediction is a measure of how close $pval(e, Y)$ is to $val(e, Y)$.
- There are many possible errors that could be measured.

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
(each Y_i is an indicator variable):

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?

Measures of error

E is the set of examples. $\mathbf{O}$ is the set of output features.

- absolute error

$$\sum_{e \in E} \sum_{Y \in \mathbf{O}} |val(e, Y) - pval(e, Y)|$$

Measures of error

E is the set of examples. $\mathbf{O}$ is the set of output features.

- absolute error

$$\sum_{e \in E} \sum_{Y \in \mathbf{O}} |val(e, Y) - pval(e, Y)|$$

- sum of squares error

$$\sum_{e \in E} \sum_{Y \in \mathbf{O}} (val(e, Y) - pval(e, Y))^2$$

Measures of error

E is the set of examples. $\mathbf{O}$ is the set of output features.

- absolute error

$$\sum_{e \in E} \sum_{Y \in \mathbf{O}} |val(e, Y) - pval(e, Y)|$$

- sum of squares error

$$\sum_{e \in E} \sum_{Y \in \mathbf{O}} (val(e, Y) - pval(e, Y))^2$$

- A cost-based error takes into account costs of various errors.

Measures of error (cont.)

When output features are $\{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} \prod_{Y \in \mathbf{O}} pval(e, Y)^{val(e,Y)} (1 - pval(e, Y))^{(1 - val(e,Y))}$$

Measures of error (cont.)

When output features are $\{0,1\}$:

- likelihood of the data

$$\prod_{e \in E} \prod_{Y \in \mathbf{O}} pval(e,Y)^{val(e,Y)} (1 - pval(e,Y))^{(1 - val(e,Y))}$$

- entropy

$$- \sum_{e \in E} \sum_{Y \in \mathbf{O}} [val(e,Y) \log pval(e,Y) + (1 - val(e,Y)) \log (1 - pval(e,Y))]$$

Point Estimates

Suppose where is a single numerical feature, Y. Let E be the training examples.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.

Point Estimates

Suppose where is a single numerical feature, Y. Let E be the training examples.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The value that minimizes the absolute error is the median value of Y.

Point Estimates

Suppose where is a single numerical feature, Y. Let E be the training examples.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The value that minimizes the absolute error is the median value of Y.
- When Y has domain $\{0, 1\}$, the prediction that maximizes the likelihood is the empirical probability.

Point Estimates

Suppose where is a single numerical feature, Y. Let E be the training examples.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The value that minimizes the absolute error is the median value of Y.
- When Y has domain $\{0, 1\}$, the prediction that maximizes the likelihood is the empirical probability.
- When Y has domain $\{0, 1\}$, the prediction that minimizes the entropy is the empirical probability.

Point Estimates

Suppose where is a single numerical feature, Y. Let E be the training examples.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The value that minimizes the absolute error is the median value of Y.
- When Y has domain $\{0, 1\}$, the prediction that maximizes the likelihood is the empirical probability.
- When Y has domain $\{0, 1\}$, the prediction that minimizes the entropy is the empirical probability.

But that doesn't mean that these predictions minimize the error for future predictions.

Training and Test Sets

To evaluate how well a learner will work on future predictions, we divide the examples into:

- training examples that are used to train the learner
- test examples that are used to evaluate the learner

...these must be kept separate.

Learning Probabilities

- Empirical probabilities do not make good predictors when evaluated by likelihood or entropy.
- Why?

Learning Probabilities

- Empirical probabilities do not make good predictors when evaluated by likelihood or entropy.
- Why? A probability of zero means "impossible" and has infinite cost.

Learning Probabilities

- Empirical probabilities do not make good predictors when evaluated by likelihood or entropy.
- Why? A probability of zero means "impossible" and has infinite cost.
- Solution: add (non-negative) pseudo-counts to the data. Suppose n_i is the number of examples with $X = v_i$, and c_i is the pseudo-count:

$$P(X = v_i) = \frac{c_i + n_i}{\sum_{i'} c_{i'} + n_{i'}}$$

- Pseudo-counts convey prior knowledge. Consider: "how much more would I believe v_i if I had seen one example with v_i true than if I has seen no examples with v_i true?"