Estimating accuracy of hypotheses

Relevant Readings: 5.1, 5.2 in Mitchell

CS495 - Machine Learning, Fall 2009

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant
  - Maybe this happened by "bad luck" and B is actually better in general

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant
  - Maybe this happened by "bad luck" and B is actually better in general
- What if the testing set had 100,000 instances?

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant
  - Maybe this happened by "bad luck" and B is actually better in general
- What if the testing set had 100,000 instances?
  - Then we have high confidence that A is more accurate than B for the measured task

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant
  - Maybe this happened by "bad luck" and B is actually better in general
- What if the testing set had 100,000 instances?
  - Then we have high confidence that A is more accurate than B for the measured task
- The is an issue of *variance in the estimate*. How do we quantify all of this?

How much can we trust the measured accuracy?

- Algorithm A an accuracy of 100% on the testing set
- Algorithm B an accuracy of 90% on the testing set
- Question: Which algorithm is better?
- Answer: We can't really say without more information.
- What if the testing set had only 10 instances?
  - The difference of one data point may not be significant
  - Maybe this happened by "bad luck" and B is actually better in general
- What if the testing set had 100,000 instances?
  - Then we have high confidence that A is more accurate than B for the measured task
- The is an issue of *variance in the estimate*. How do we quantify all of this?

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$
- The error rate that some hypothesis makes on S is called the *sample error*

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$
- The error rate that some hypothesis makes on S is called the *sample error*
  - Easy to measure; just use the testing set

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$
- The error rate that some hypothesis makes on S is called the *sample error*
  - Easy to measure; just use the testing set
- The error rate that some hypothesis makes on X (under distribution $\mathcal{D}$) is called the *true error*

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$
- The error rate that some hypothesis makes on S is called the *sample error*
  - Easy to measure; just use the testing set
- The error rate that some hypothesis makes on X (under distribution $\mathcal{D}$) is called the *true error*
  - Harder to measure, but more important to know

The skiing example [Mitchell, 5.2]

- The instance space X is the set of all people
  - This includes some attributes (age, time since last visit, etc.)
- There is a probability distribution $\mathcal{D}$ (unknown to us) that gives the probability that person $x \in X$ is the next one to walk in the ski shop
- Target concept $f : X \to \{0, 1\}$ is whether or not each person plans to buy skis when they visit
- Suppose sample S is drawn from X according to $\mathcal{D}$
- The error rate that some hypothesis makes on S is called the *sample error*
  - Easy to measure; just use the testing set
- The error rate that some hypothesis makes on X (under distribution $\mathcal{D}$) is called the *true error*
  - Harder to measure, but more important to know

Estimating true error (confidence intervals)

- The sample error provides an estimate of true error, but how good is the estimate?

Estimating true error (confidence intervals)

- The sample error provides an estimate of true error, but how good is the estimate?
- Under some reasonable assumptions (see [Mitchell, 5.2.2]), we can say with 99% probability that:
  - $e - 2.58\sqrt{e(1-e)/n} \leq E \leq e + 2.58\sqrt{e(1-e)/n}$
  - where e is the sample error,
  - E is the true error, and
  - $n \geq 30$ is the number of samples

Estimating true error (confidence intervals)

- The sample error provides an estimate of true error, but how good is the estimate?
- Under some reasonable assumptions (see [Mitchell, 5.2.2]), we can say with 99% probability that:
  - $e - 2.58\sqrt{e(1-e)/n} \leq E \leq e + 2.58\sqrt{e(1-e)/n}$
  - where e is the sample error,
  - E is the true error, and
  - $n \geq 30$ is the number of samples
- The 2.58 constant can be adjusted according to the certainty we require:

Estimating true error (confidence intervals)

- The sample error provides an estimate of true error, but how good is the estimate?
- Under some reasonable assumptions (see [Mitchell, 5.2.2]), we can say with 99% probability that:
  - $e - 2.58\sqrt{e(1-e)/n} \leq E \leq e + 2.58\sqrt{e(1-e)/n}$
  - where e is the sample error,
  - E is the true error, and
  - $n \geq 30$ is the number of samples
- The 2.58 constant can be adjusted according to the certainty we require:
  - For 99% certainty use 2.58
  - For 98% certainty use 2.33
  - For 95% certainty use 1.96
  - For 90% certainty use 1.64
  - For 80% certainty use 1.28
  - For 68% certainty use 1.00
  - For 50% certainty use 0.67

Estimating true error (confidence intervals)

- The sample error provides an estimate of true error, but how good is the estimate?
- Under some reasonable assumptions (see [Mitchell, 5.2.2]), we can say with 99% probability that:
  - $e - 2.58\sqrt{e(1-e)/n} \leq E \leq e + 2.58\sqrt{e(1-e)/n}$
  - where e is the sample error,
  - E is the true error, and
  - $n \geq 30$ is the number of samples
- The 2.58 constant can be adjusted according to the certainty we require:
  - For 99% certainty use 2.58
  - For 98% certainty use 2.33
  - For 95% certainty use 1.96
  - For 90% certainty use 1.64
  - For 80% certainty use 1.28
  - For 68% certainty use 1.00
  - For 50% certainty use 0.67