CS209: Machine Learning

Hypothesis

Sachchida Nand Chaurasia

Assistant Professor

Department of Computer Science
Banaras Hindu University
Varanasi
Email id: snchaurasia@bhu.ac.in, sachchidanand.mca07@gmail.com

Outline

Hypothesis Testing

Types of errors

Bias & Variance, Best fitting, Overfitting and Underfitting

Outline

Hypothesis Testing

Types of errors

Bias & Variance, Best fitting, Overfitting and Underfitting

Hypothesis Test I

Conduct and interpret results from a hypothesis test about a population mean.

- Recognize when to use a hypothesis test or a confidence interval to draw a conclusion about a population mean.
- Under appropriate conditions, conduct a hypothesis test about a population mean. State a conclusion in context.

Hypothesis Test II

Inference for Means: Focus is on inference when the variable is quantitative. Here parameters and statistics are means.

Estimating a Population Mean: Learned how to use a sample mean to calculate a confidence interval. The confidence interval estimates a population mean.

✓ Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population.

✓ Hypothesis testing uses sample data to evaluate a hypothesis about a population.

✓ The first step in hypothesis test is setting of hypothesis.

1. Null hypothesis: It is a claim or statement about population parameter that is assumed to be true until it declared to be false. It is denoted by $\mathbf{H}_0$

Hypothesis Test III

2. Alternative hypothesis (Research hypothesis): Any hypothesis which is complementary to null hypothesis. It is denoted by $\mathbf{H}_1$.

Example:

$$\mathbf{H}_0\text{: } \mu = \text{x}, \ \mathbf{H}_1\text{: } \mu \neq \text{x}$$
$$\mathbf{H}_0\text{: } \mu \leq \text{x} \ \mathbf{H}_1\text{: } \mu > \text{x}$$
$$\mathbf{H}_0\text{: } \mu \geq \text{x}, \ \mathbf{H}_1\text{: } \mu < \text{x}$$

✓ $\mathbf{H}_0$ is to make a decision (yes or no)

✓ The alternative hypothesis give an answer for a research question about validity of claim. For example: <1000ml or >1000ml.

✓ H_0 **and** H_1 **must be mutually exclusive, and** H_1 **should not contain equality**

Hypothesis Test IV

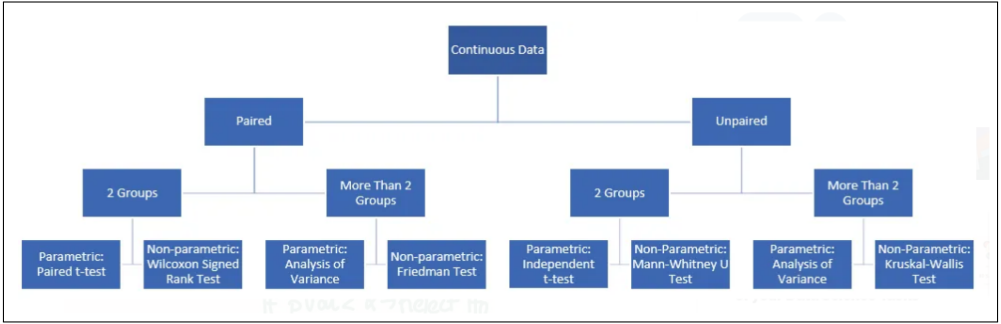

Continuous Data
Paired
Unpaired
2 Groups
More Than 2 Groups
2 Groups
More Than 2 Groups
Parametric: Paired t-test
Non-parametric: Wilcoxon Signed Rank Test
Parametric: Analysis of Variance
Non-parametric: Friedman Test
Parametric: Independent t-test
Non-Parametric: Mann-Whitney U Test
Parametric: Analysis of Variance
Non-Parametric: Kruskal-Wallis Test

Figure 1: Test Methods

Hypothesis Test V

Computation of test statistic:

Test of significance:

z-test, t-test, Chi-test, f-test are some frequently used testing method.

z-test:

n$\geq$ 30, σ =S.D. of population is known, then

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} \sim \mathcal{N}(0, 1) \tag{1}$$

$\sigma_{\bar{X}}$ = Standard error of mean = $\frac{\sigma}{\sqrt{n}}$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} \sim \mathcal{N}(0, 1) \tag{2}$$

Hypothesis Test VI

$$\sigma_{\bar{X}} = \text{Standard error of mean} = \frac{S_\sigma}{\sqrt{n}},\ S_\sigma \rightarrow \text{Sample S.D.}$$

Q: A Telecom service provider claims that individual customers pay on an average 400 rs. per month with standard deviation of 25 rs. A random sample of 50 customers bills during a given month is taken with a mean of 250 and standard deviation of 15. What to say with respect to the claim made by the service provider?

Solution:

First thing first, Note down what is given in the question:

H_0 (Null Hypothesis) : $\mu = 400$

H_1 (Alternate Hypothesis): $\mu \neq 400$ (Not equal means either $\mu > 400$ or $\mu < 400$. Hence it will be validated with two tailed test)

$\sigma = 25$ (Population Standard Deviation)

Level of Significance (LoS) (α) = 5% (Take 5% if not given in question)

Hypothesis Test VII

n = 50 (Sample size)

$\bar{x}$= 250 (Sample mean)

s = 15 (sample Standard deviation)

n $\geq$ 30 hence will go with z-test

Step 1:

Calculate z using z-test formula as below:

z = ($\bar{x}$ -μ)/ (σ/$\sqrt{n}$)

z = (250 - 400) / (25/$sqrt50$)

z = -42.42

Step 2:

get z critical value from z table for α = 5$\%$

z critical values = (-1.96, +1.96)

Hypothesis Test VIII

to accept the claim (significantly), calculated z should be in between

-1.96 < z < +1.96

but calculated z (-42.42) < -1.96 which mean reject the null hypothesis

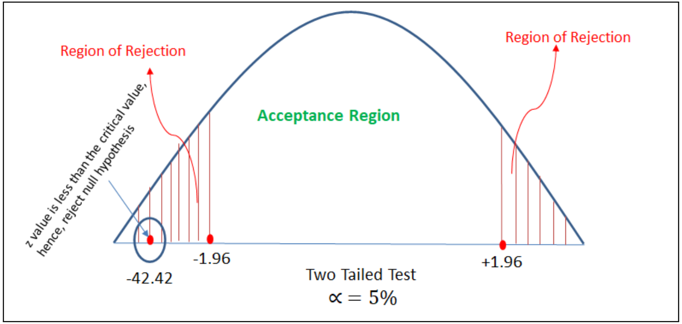


Figure 2: z-test

Hypothesis Test IX

Q: From the data available, it is observed that 400 out of 850 customers purchased the groceries online. Can we say that most of the customers are moving towards online shopping even for groceries?

Hypothesis Test X

Solution:

Note down what is given:

400 out of 850 which indicates that this is a proportion problem.

Proportion p (small p) = 400/850 = 0.47

H_0 (Null Hypothesis): P (capital P) > 0.5 (claim is that most of the customers are moving towards online shopping even for groceries which mean at least 50% should do online shopping)

H_1 (Alternate Hypothesis): P $\leq$ 0.5 left tailed

n = 850

LoS (α) = 5% (assume 5% as it not given in question)

n > = 30 hence will go with z-test

Step 1:

calculate z value using the z-test formula

Hypothesis Test XI

z = (p - P)/$\sqrt{(P * Q/n)}$

z = (0.47 - 0.50)/$\sqrt{(0.5 * 0.5/850)}$

z = 1.74

Step 2:

get z value from z table for α = 5%

From z-table, for α = 5%, z = -1.645 (one value as it is one (left) tailed problem)

z(calculated) 1.74 > -1.645 (z from z-table with α = 5%)

Conclusion:

Hence accept the null hypothesis that mean, with given data we can validate significantly that most of the customers are moving towards online shopping even for groceries.

Hypothesis Test XII

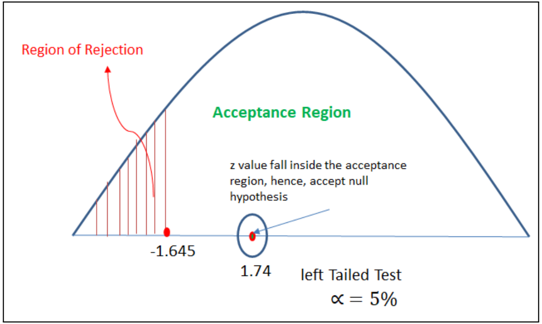

Region of Rejection
Acceptance Region
z value fall inside the acceptance region, hence, accept null hypothesis
-1.645
1.74
left Tailed Test
∝ = 5%

Figure 3: z-test

Hypothesis Test XIII

Q: It is found that 250 errors in the randomly selected 1000 lines of code from Team A and 300 errors in 800 lines of code from Team B. Can we assume that team B's performance is superior to that of A.

Hypothesis Test XIV

Solution:

Note down what is given in the question:

There are two samples : Team A and Team B

for each Team some proportion is given in terms of line of error out of total line of code.

Hence this problem can be solved using two proportion z-test.

For one or two proportion type problem we use z-test. (in case of multi-proportion we use $x2$ that is chi square test)

Team A (Sample A):

proportion pA (small p) = 250/1000 = 0.25

nA = 1000

Team B (Sample B):

proportion pB (small p) = 300/800 = 0.375

Hypothesis Test XV

nB = 800

Take α = 5% (assume α = 5% if not given in question)

Claim: Team B's performance is superior than Team A which means:

H_0 (Null Hypothesis): overall mean error of Team B μB < μA overall mean error of Team A (with respect to population)

H_1 (Alternate Hypothesis) : μB > = μA (one right tailed test)

Step 1:

calculate z value from two proportion z-test formula as below:

z = (pA - pB)/[p^(1-p^)(1/nA + 1/nB)]

where p^ (p hat) = (nA*pA + nB*pB)/(nA + nB)

p^ (p hat) = (1000*0.25 + 800*0.375) / (1000 + 800) = 0.305

z = (0.25 - 0.375) / [0.305*(1-0.305)*(1/1000 + 1/800)]

Hypothesis Test XVI

z = -0.125/[0.212*0.00135]

z = -436.75

Step 2:

get z using z-table for α = 5% which is z = +1.645

Now calculated z -436.75 < +1.645

Hence will conclude that null hypothesis is true which mean from given data it is proven significantly that team B's performance is better that team A's performance.

Hypothesis Test XVII

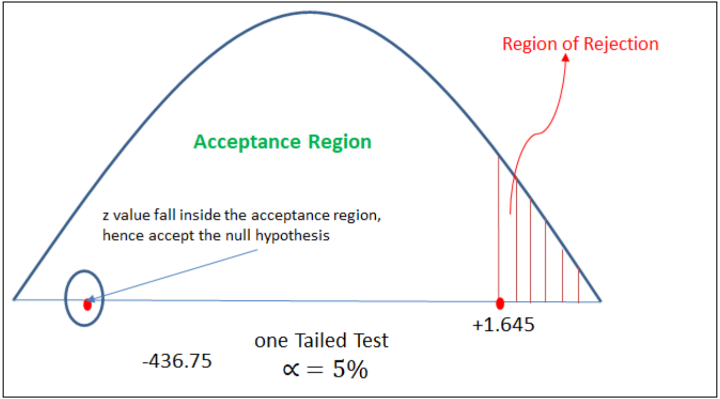


Figure 4: z-test

Hypothesis Test XVIII

Q: Following is the record of number of accidents took place during the various days of the week.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
120	140	200	90	140	120	180

Accidents took place in various days of the given week.

Can we conclude that accidents are independent of the day of week?

Solution:

H_0 (Null Hypothesis): Accidents are independent of the day

H_1 (Alternate Hypothesis): Not independent

Here each day will represent a sample and observed accident data is proportion.

Hence this problem can be categorized as multi proportion problem and will be solved using χ 2 chi square test.

Hypothesis Test XIX

Below table shows the Observed values of accident in first column.

As we want to validate that accidents are independent of the day of week.

Observed (o)	Expected (e = average of Observed values)	χ 2 = $\sum$[(o-e)2]/e
120	110	0.909
140	110	8.181
200	110	73.636
90	110	3.636
140	110	8.181
120	110	0.909
180	110	44.545
Total = 990		Total χ 2 = 139.997

Hypothesis Test XX

for that average accidents on each day should be different.

Hence we need to calculate average accidents on each day and this will be called as expected value in χ 2 test.

Take α = 5% (assume α = 5% if not given in question)

Step 1: calculate expected values and χ 2 values using χ 2 formula as shown below in the table.

Step 2: use χ 2 table for α = 5% and get χ 2 value from the table.

from table we got χ 2 (critical value at α = 5%) = 3.841

Step 3: compare both χ 2 values.

The chi-square value of 18.99 is much larger than the critical value of 3.84, so the null hypothesis can be rejected.

Hypothesis Test XXI

It means, reject the null hypothesis and accept the alternate hypothesis. Which means with given data we can conclude significantly that accidents are not independent of the day of week. [might not look realistic but with given data is concluding this]

Hypothesis Test XXII

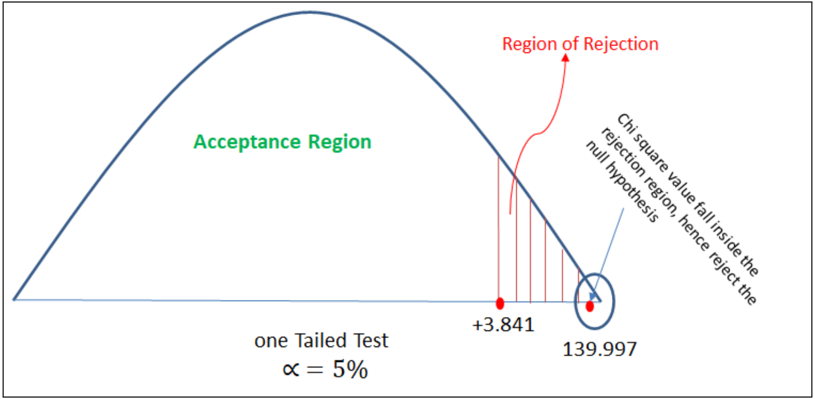

Region of Rejection
Acceptance Region
Chi square value fall inside the rejection region, hence reject the null hypothesis
+3.841
139.997
one Tailed Test
$\propto = 5\%$

Figure 5: χ-test

Outline

Hypothesis Testing

Types of errors

Bias & Variance, Best fitting, Overfitting and Underfitting

Type I and Type II errors I

✓ In statistical hypothesis testing, a type I error is the mistaken rejection of the null hypothesis (also known as a "false positive" finding or conclusion; **Example: "an innocent person is convicted"**)

✓ While a type II error is the mistaken acceptance of the null hypothesis (also known as a "false negative" finding or conclusion; **Example: "a guilty person is not convicted"**).

✓ When we are testing Null Hypothesis(H_0) against Alternative Hypothesis(H_1) there are four possibilities:

1. H_0 accepted when H_0 is true [Correct]
2. H_0 rejected when H_0 is true [Type I error]
3. H_0 accepted when H_0 is false [Type II error]
4. H_0 rejected when H_0 is false [Correct]

Type I and Type II errors II

Probability of types of errors:

Type I error	Type II error
α= p(Type I error)	β=p(Type II error)
α = p(reject $H_0 \| H_1$ is true)	β=p(accept $H_0 \| H_1$ is true)

Outline

Hypothesis Testing

Types of errors

Bias & Variance, Best fitting, Overfitting and Underfitting

Bias and Variance I

✓ It is important to understand prediction errors (bias and variance) when it comes to accuracy in any machine learning algorithm.

✓ There is a tradeoff between a model's ability to minimize bias and variance which is referred to as the best solution for selecting a value of Regularization constant.

✓ A proper understanding of these errors would help to avoid the overfitting and underfitting of a data set while training the algorithm.

Bias:

✓ The bias is known as the difference between the prediction of the values by the ML model and the correct value.

✓ Being high in biasing gives a large error in training as well as testing data.

✓ Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.

Bias and Variance II

✓ By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set.

✓ Such fitting is known as Underfitting of Data. This happens when the hypothesis is too simple or linear in nature.

Variance:

✓ The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model.

✓ The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn't seen before.

✓ As a result, such models perform very well on training data but has high error rates on test data.

✓ When a model is high on variance, it is then said to as Overfitting of Data.

Bias and Variance III

✓ Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.

Tick While training a data model variance should be kept low.

Bias and variance tradeoff:

✓ In statistics and machine learning, the bias–variance trade-off is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

✓ The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.

✓ The bias error is an error from erroneous assumptions in the learning algorithm.

✓ *Unfortunately, it is typically impossible to do both simultaneously*.

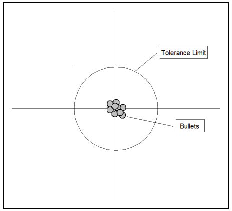
Bias and Variance IV

High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting)

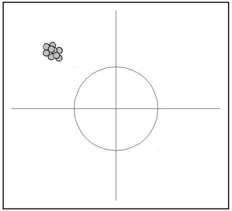
The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting)

✓ The bias–variance decomposition is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

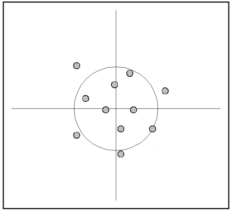
Bias and Variance V



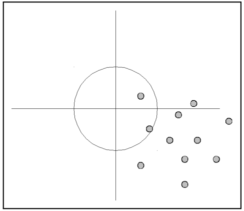
(a) low bias and low variance



(b) high bias and low variance



(c) low bias and high



(d) high bias and high

Bias and Variance VI

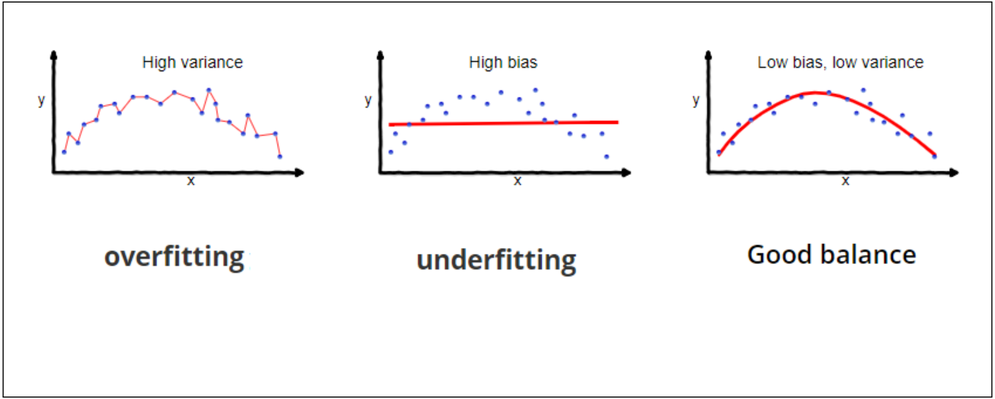


Figure 6: Fitting

Bias and Variance VII

✓ An epoch is a term used in machine learning and indicates the number of passes of the entire training dataset the machine learning algorithm has completed. Datasets are usually grouped into batches (especially when the amount of data is very large).

Bias and Variance VIII

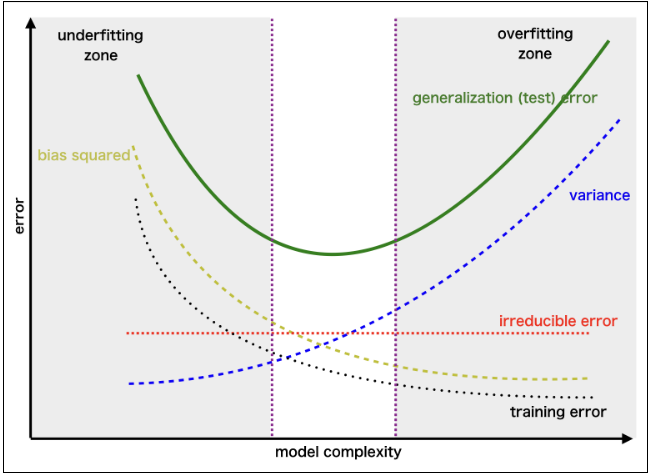


Figure 7: Bias and variance as function of model complexity

Bias and Variance IX

Mathematically

Suppose that we have a training set consisting of a set of points $x_1, \dots, x_n$ and real values y_i associated with each point x_i. We assume that there is a function

$$Y = f(X) + e \tag{3}$$

where the noise, e, has zero mean and variance σ^2.

✓ Let the variable we are trying to predict as Y and other covariates as X. We assume there is a relationship between the two such that

So the expected squared error at a point x is

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right] \tag{4}$$

Bias and Variance X

The Err(x) can be further decomposed as

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2 \tag{5}$$

$$Err(x) = Bias^2 + Variance + IrreducibleError \tag{6}$$

✓ Err(x) is the sum of Bias^2, variance and the irreducible error.

Irreducible error is the error that can't be reduced by creating good models. It is a measure of the amount of noise in our data. Here it is important to understand that no matter how good we make our model, our data will have certain amount of noise or irreducible error that can not be removed.

Bias and Variance XI

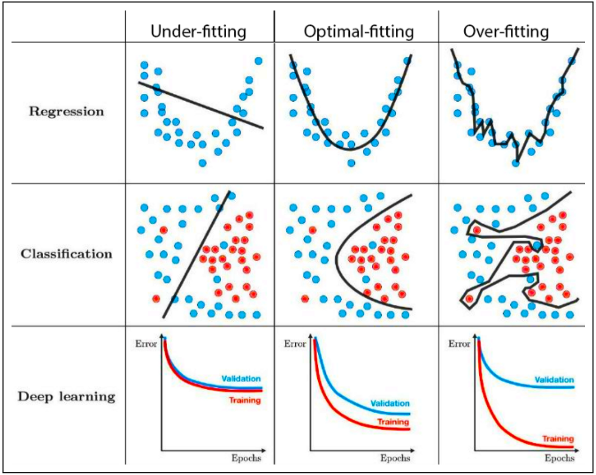


Figure 8: Fitting

Bias and Variance XII

Approaches:

✓ Dimensionality reduction and feature selection can decrease variance by simplifying models. Similarly, a larger training set tends to decrease variance.

✓ Adding features (predictors) tends to decrease bias, at the expense of introducing additional variance. Learning algorithms typically have some tunable parameters that control bias and variance; for example,

- linear and Generalized linear models can be regularized to decrease their variance at the cost of increasing their bias.
- In artificial neural networks, the variance increases and the bias decreases as the number of hidden units increase, although this classical assumption has been the subject of recent debate. Like in GLMs, regularization is typically applied.

Bias and Variance XIII

- In k-nearest neighbor models, a high value of k leads to high bias and low variance (see below).
- In instance-based learning, regularization can be achieved varying the mixture of prototypes and exemplars.
- In decision trees, the depth of the tree determines the variance. Decision trees are commonly pruned to control variance.

✓ One way of resolving the trade-off is to use mixture models and ensemble learning.

✓ For example, boosting combines many "weak" (high bias) models in an ensemble that has lower bias than the individual models, while bagging combines "strong" learners in a way that reduces their variance.

✓ Model validation methods such as cross-validation (statistics) can be used to tune models so as to optimize the trade-off.

Applications:

In regression: The bias–variance decomposition forms the conceptual basis for regression regularization methods such as Lasso and ridge regression.

In classification: The bias–variance decomposition was originally formulated for least-squares regression. For the case of classification under the 0-1 loss (misclassification rate), it is possible to find a similar decomposition.

In reinforcement learning: Even though the bias–variance decomposition does not directly apply in reinforcement learning, a similar tradeoff can also characterize generalization. When an agent has limited information on its environment, the suboptimality of an RL algorithm can be decomposed into the sum of two terms: a term related to an asymptotic bias and a term due to overfitting.

Bias and Variance XV

In human learning: While widely discussed in the context of machine learning, the bias–variance dilemma has been examined in the context of human cognition, most notably by Gerd Gigerenzer and co-workers in the context of learned heuristics.