Что такое смещение?

Смещение — это разница между средним прогнозом нашей модели и правильным значением, которое мы пытаемся предсказать. Модель с высоким смещением очень мало внимания уделяет обучающим данным и чрезмерно упрощает модель. Это всегда приводит к высокой ошибке на обучающих и тестовых данных.

Что такое дисперсия?

Дисперсия — это изменчивость предсказания модели для данной точки данных или значения, которое говорит нам о разбросе наших данных. Модель с высокой дисперсией уделяет большое внимание обучающим данным и не обобщает данные, которые она раньше не видела. В результате такие модели очень хорошо работают на обучающих данных, но имеют высокий уровень ошибок на тестовых данных.

Математически

Пусть переменная, которую мы пытаемся предсказать, представляет собой Y, а другие ковариаты — как X. Мы предполагаем, что между ними существует взаимосвязь, такая, что

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

Где e — член ошибки, и он обычно распределяется со средним значением 0.

Мы создадим модель f ^ (X) для f (X), используя линейную регрессию или любой другой метод моделирования.

Таким образом, ожидаемая квадратичная ошибка в точке x равна

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

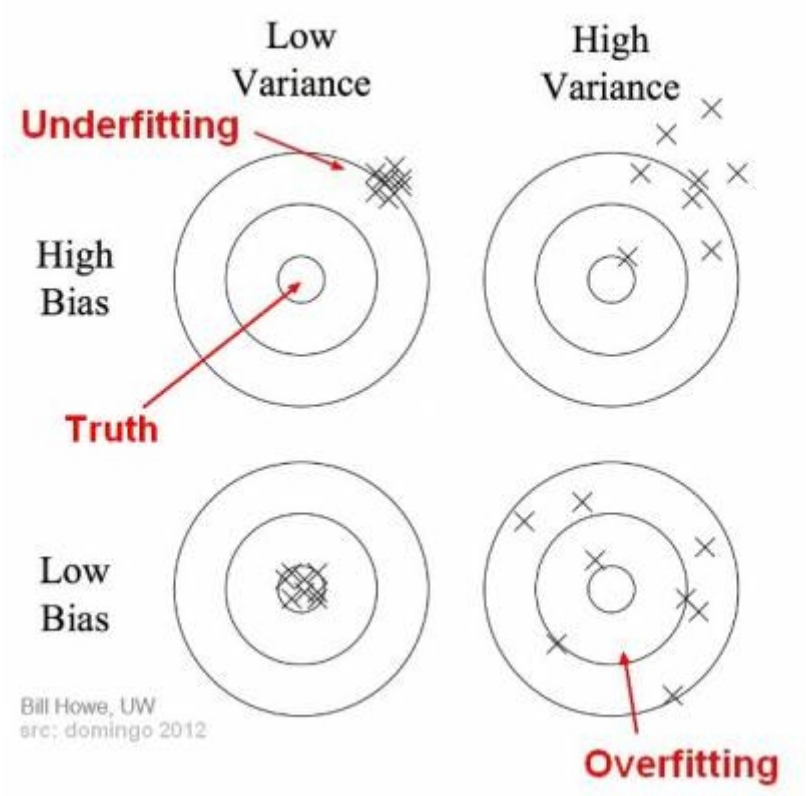
Err(x) можно далее разложить как

$$Err(x)=\mathrm{Bias}^2+\mathrm{Variance}+\mathrm{Irreducible\ Error}$$

Err(x) — это сумма смещение², дисперсии и неустранимой ошибки.

Неустранимая ошибка — это ошибка, которую нельзя уменьшить, создавая хорошие модели. Это мера количества шума в наших данных. Здесь важно понимать, что как бы хорошо мы ни сделали нашу модель, наши данные будут иметь определенный уровень шума или неустранимую ошибку, которую невозможно будет удалить.

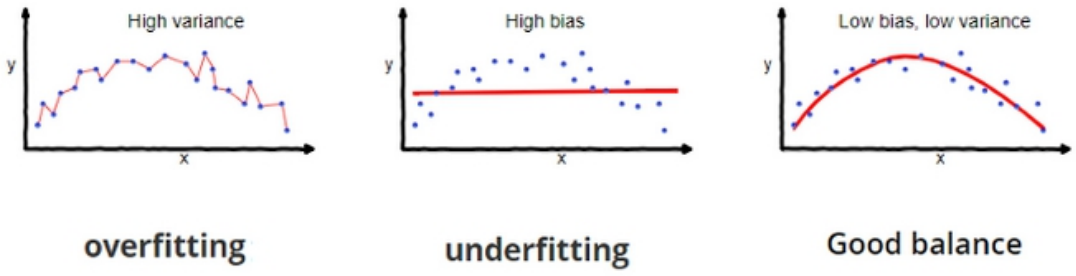
Смещение и дисперсия с использованием диаграммы



На приведенной выше диаграмме центр цели — это модель, которая точно предсказывает правильные значения. По мере того, как мы удаляемся от мишени, наши прогнозы становятся все хуже и хуже. Мы можем повторить наш процесс построения модели, чтобы получить отдельные попадания в цель.

В обучении с учителем недообучение происходит, когда модель не может уловить базовую структуру данных. Эти модели обычно имеют высокое смещение и низкую дисперсию. Это происходит, когда у нас очень мало данных для построения точной модели или когда мы пытаемся построить линейную модель с нелинейными данными. Кроме того, такие модели очень просты для захвата сложных закономерностей в данных, таких как линейная и логистическая регрессия.

В обучении с учителем переобучение происходит, когда наша модель фиксирует шум вместе с лежащим в основе шаблоном в данных. Это происходит, когда мы много тренируем нашу модель на зашумленном наборе данных. Эти модели имеют низкое смещение и высокую дисперсию. Эти модели очень сложны, как и деревья решений, которые склонны к переобучению.



Почему смещение дисперсии является компромиссом?

Если наша модель слишком проста и имеет очень мало параметров, она может иметь большое смещение и низкую дисперсию. С другой стороны, если наша модель имеет большое количество параметров, она будет иметь высокую дисперсию и низкое смещение. Поэтому нам нужно найти правильный/хороший баланс без переобучения и недообучения данных.

Общая ошибка

Чтобы построить хорошую модель, нам нужно найти хороший баланс между смещением и дисперсией, чтобы минимизировать общую ошибку.

Total Error = Bias^2 + Variance + Irreducible Error