**Bias-Variance Tradeoff**

Kompromis między odchyleniem a wariancją

Nie pożądane jest nadmierne uproszczenie (wysokie odchylenie), jak i nadmiernie szczegółowa wrażliwość (wysoka wariancja).

W nadzorowanym uczeniu maszynowym algorytm uczy się modelu na podstawie danych treningowych. Celem każdego nadzorowanego algorytmu uczenia maszynowego jest jak najlepsze oszacowanie funkcji mapowania (f) dla zmiennej wyjściowej (Y), biorąc pod uwagę dane wejściowe (X). Funkcja mapowania jest często nazywana funkcją docelową, ponieważ jest to funkcja, którą dany nadzorowany algorytm uczenia maszynowego ma do przybliżenia.

Błąd prognozowania dla dowolnego algorytmu uczenia maszynowego można podzielić na trzy części:

Błąd odchylenia (Bias Error)

Błąd wariancji (Variance Error)

Błąd nieredukowalny (Irreducible Error)

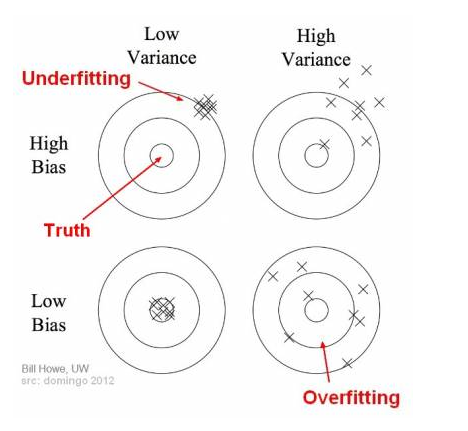
Błąd nieredukowalny nie może zostać zmniejszony bez względu na zastosowany algorytm. Jest to błąd wprowadzony w ramach wybranego problemu i może być spowodowany czynnikami takimi jak nieznane zmienne, które wpływają na mapowanie zmiennych wejściowych na zmienną wyjściową.

**Bias** - jest różnicą między średnią prognozą naszego modelu a prawidłową wartością, którą próbujemy przewidzieć. Model o dużym odchyleniu przywiązuje bardzo małą wagę do danych treningowych i **upraszcza model**. Zawsze prowadzi to do wysokiego błędu w danych treningowych i testowych. Działa słabo na zestawie szkoleniowym i testowym, ponieważ nie jest w stanie zidentyfikować wzorców w danych. Wskaźniki oceny, takie jak dokładność, wynik F1 dla takich modeli, które cierpią na wysokie odchylenie, jest bardzo niski, ponieważ różnica między wartościami przewidywanymi i rzeczywistymi jest duża.

**Variance** - zmienność prognoz modelu dla danego punktu danych lub wartości, która mówi nam o rozprzestrzenianiu się naszych danych. Model o dużej wariancji przywiązuje dużą wagę do danych treningowych i nie uogólnia na dane, których wcześniej nie widział. W rezultacie takie modele działają bardzo dobrze na danych szkoleniowych, ale mają wysoki poziom błędów w danych testowych. Innymi słowy - Wariancja odnosi się do zdolności modelu do pomiaru rozprzestrzeniania się danych. Duża wariancja lub **nadmierne dopasowanie** oznacza, że model pasuje do dostępnych danych, ale nie uogólnia się dobrze, aby przewidzieć nowe dane.

**High-Bias** **and Low-Variance** machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

**Low-Bias and High-Variance** machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines



Na powyższym diagramie środkiem celu jest model, który doskonale przewiduje prawidłowe wartości. W miarę oddalania się od byczego oka nasze przewidywania stają się coraz gorsze.

**Niedopasowanie (Underfitting)** występuje, gdy model nie jest w stanie uchwycić podstawowego wzorca danych. Modele te mają zwykle dużą stronniczość i niską wariancję (**High-Bias** **and Low-Variance)**. Dzieje się tak, gdy mamy bardzo mało danych do zbudowania dokładnego modelu lub gdy próbujemy zbudować model liniowy z danymi nieliniowymi.

**Nadmierne dopasowanie** **(Overfitting)** ma miejsce, gdy nasz model wychwytuje hałas wraz z podstawowym wzorcem danych. Zdarza się to, gdy dużo trenujemy nasz model w hałaśliwym zestawie danych. Modele te mają niskie odchylenie i dużą wariancję( **Low-Bias and High-Variance)**. Modele te są bardzo złożone, podobnie jak drzewa decyzyjne, które są podatne na nadmierne dopasowanie.

