1. Problemy pojawiające się w praktyce w zadaniu uczenia maszynowego: przeuczenie (ang. overfitting), klątwa  
   wymiarowości, występowanie wartości odstających. Przykładowe sposoby zmniejszenia wpływu każdego z  
   tych zjawisk.

*Piotr Laskowski*

**Przeuczenie (overfitting):**

Przeuczenie modelu następuje wtedy, gdy wyprodukowany model jest zbyt ciasno dopasowany do danych na których jest trenowany. Dzieje się tak dlatego, że szum powiązany z nadmierną ilością losowych danych w zbiorze uczącym składają się na sytuację w której model staje się zbyt skonkretyzowany i nastawiony na detale parametrów uczących występujące jedynie w konkretnym zbiorze danych. To z kolei skutkuje doskonałym dopasowaniem modelu do danych testowych powstających jako derywat naszego zbioru głównego, natomiast wpływa negatywnie na dopasowanie do świeżych danych. W wyniku tego wytworzony model jest niejako “jednokrotnego użytku” i traci możliwości trafnej generalizacji wyuczonych schematów.

Przykładowe sposoby zmniejszenia wpływu tego zjawiska:

1. **Sprawdzian krzyżowy (Cross-validation)** - wygenerowanie mnogości mniejszych próbek testowo treningowych do odpowiedniego dostosowania hiperparametrów uczących modelu. Dzięki takiemu podziałowi zestaw testowy jest odseparowany i stanowi niejako zupełnie nowy zestaw danych dla modelu
2. **Wykorzystanie większej ilości danych** - użycie większej ilości danych powinno spowodować ulepszony poziom detekcji trafnych sygnałów przez algorytm. Jest to jednak pomocne tylko w przypadku użycia danych pozbawionych nadmiernej ilości szumów
3. **Selekcja cech** - w przypadku parametrów opornych na generalizację pomocne może okazać się ograniczenie ich ilości.
4. **Ograniczenie liczby iteracji treningowych** - krzywa iteracyjna ma tendencję do wzrostu wraz z ilością powtórzeń aż do pewnego punktu. Kiedy go osiąga spada umiejętność trafnej generalizacji danych przez model co powoduje nadmierne dopasowanie

**Klątwa wymiarowości (curse of dimensionality):**

Zjawisko objawiające się spadkiem gęstości danych wraz ze wzrostem ich wymiarowości występujące przy wykorzystaniu klasyfikatorów. Istnieje bowiem optymalna liczba cech opisujących dane wykorzystywane w modelu odznaczająca się najwydajniejszym dopasowaniem przez klasyfikator.

Przykładowe sposoby zmniejszenia wpływu tego zjawiska:

1. **Manualna selekcja cech** - ograniczenie liczby parametrów branych pod uwagę przy uczeniu modelu (tym samym wymiarowości) do tych, które odznaczają się najwyższym poziomem korelacji (np. Pearsona)  ze zmienną decyzyjną
2. **PCA** - wykorzystanie algorytmu analizy zmiennych składowych - faktoryzacja cech
3. **Skalowanie wielowymiarowe** - wykorzystanie algorytmu zachowującego strukturę metryczną obiektów w naszym układzie a zarazem mapujących te odległości w przestrzeni o mniejszej ilości wymiarów

**Występowanie wartości odstających (outliers):**

Zjawisko występowania wartości odstających polega na identyfikacji obserwacji, które są relatywnie odległe od pozostałych elementów próbki danych. Oznacza to w dużym uproszczeniu, że związek pomiędzy taką obserwacją a zmienną decyzyjną wpływa niekorzystnie na jej predykcję pod kątem poziomu istotności w kontekście danych wejściowych.

Przykładowe sposoby zmniejszenia wpływu tego zjawiska:

1. Usunięcie redundantnych obserwacji - podczas gdy są one wynikiem błędnie wprowadzonych danych lub błędu podczas preprocesowania danych wejściowych zakładając, że ich liczebność jest niewielka
2. Transformacja wartości parametrów obserwacji odstających - normalizacja argumentów w taki sposób, aby zredukować wpływ odstającego charakteru ich wartości. Odbywa się to najczęściej z wykorzystaniem metod skalujących lub transformacji logarytmicznych
3. Imputacja wartości odstających - tak jak w przypadku brakujących danych, możemy wygenerować na podstawie analizy podstawowych atrybutów statystycznych modelu takie wartości dla cech posiadających tendencję do odstawania, aby nie utracić na spójności modelu - wykorzystując np. średnią. Dzięki zastosowaniu tej techniki nie tracimy danych.