Metody Inżynierii Wiedzy
Informacja, dane i wiedza - modele parametryczne i nieparametryczne - wykład 2

Adam Szmigielski
aszmigie@pjwstk.edu.pl
materiały: $ftp(public) : //aszmigie/MIW$

Prawdopodobieństwo a informacja

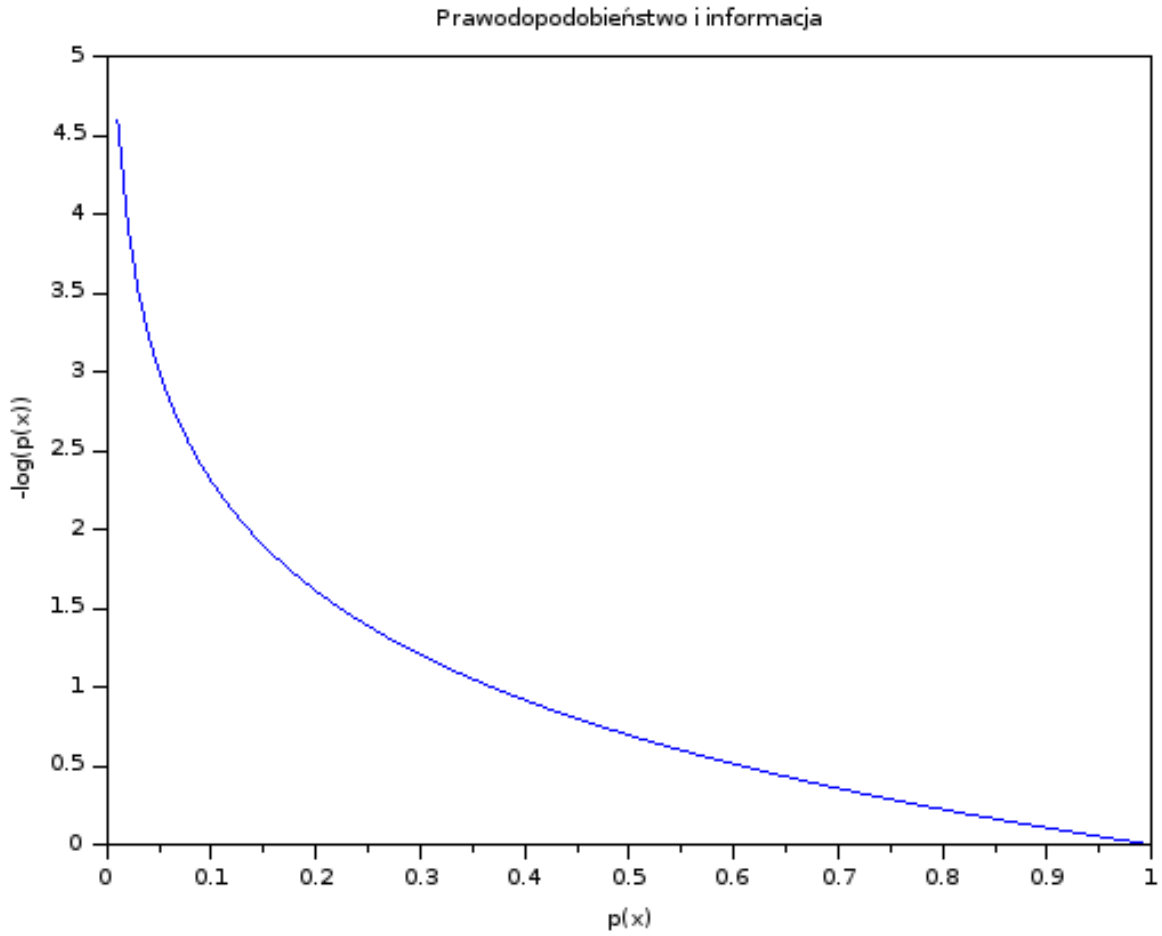
- Ile jest informacji, gdy obserwujemy konkretną wartość zmiennej losowej x ?
- Ta ilość informacji może być postrzegana jako "stopień zaskoczenia" w procesie obserwacji wartości x:
  - mało prawdopodobne obserwacja x - dużo informacji,
  - bardzo prawdopodobna obserwacja - mało informacja.
- Miara zawartości informacji $h(x)$ będzie zależała od prawdopodobieństwa $p(x)$:

$$h(x) = -\log_2 p(x)$$

- Jeśli mamy dwa zdarzenia x i y, które są statystycznie niezależne (tj. $p(x, y) = p(x) \cdot p(y)$), informacja z obserwacji obu z nich powinna być sumą informacji uzyskanych od każdego z nich osobno tj.

$$h(x, y) = h(x) + h(y)$$

Związek między prawdopodobieństwem a informacją



Entropia Shanona - wartość oczekiwana informacji

- Załóżmy, że nadawca chce przekazać wartość zmiennej losowej.
- Ilość informacji, w odniesieniu do rozkładu prawdopodobieństwa $p(x)$, jest określona jako wartość oczekiwana informacji $h(x)$ rozkładu:

$$H[x] = \sum_x p(x) \cdot h(x) = -\sum_x p(x) \cdot \log(p(x))$$

  Wielkość ta nazywana jest **entropią zmiennej losowej** x.
- Zauważmy, że $\lim_{p \to 0}(p \cdot \log(p) = 0)$ tj. $p(x)\log(p(x)) = 0$ dla $p(x) = 0$.

Entropia - przykład

- Rozważmy zmienną losową x mającą 8 możliwych stanów $\{a, b, c, d, e, f, g, h\}$, z których każdy jest równie prawdopodobny. Entropia takiego rozkładu wynosi:

$$H[x] = -8 \times \frac{1}{8} \cdot \log_2(\frac{1}{8}) = 3$$

- Dla innych prawdopodobieństw stanów np. $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$ entropia wynosi odpowiednio:

$$H[x] = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{4}\log_2(\frac{1}{4}) - \frac{1}{8}\log_2(\frac{1}{8}) - \frac{1}{16}\log_2(\frac{1}{16}) - \frac{4}{64}\log_2(\frac{1}{64}) = 2$$

- Niejednolity rozkład ma mniejszą entropię niż jednorodną

Korelacja

Sprawdza czy jakiekolwiek dwie cechy, atrybuty lub własności (wyrażone liczbowo) współwystępują ze sobą. Obliczany współczynnik zawsze waha się od -1 do 1

- **Współczynnik korelacji** – liczba określająca w jakim stopniu zmienne są współzależne.
- Istnieje wiele różnych wzorów określanych jako współczynniki korelacji.
- Większość z nich jest znormalizowana: +1 (zupełna korelacja dodatnia), -1 (zupełna korelacja ujemna), 0 (brak korelacji),
- Najczęściej stosowany jest współczynnik korelacji r Pearsona - korelacja liniowa.

Współczynnik korelacji liniowej Pearsona

- *Współczynnik korelacji liniowej Pearsona* - współczynnik określający poziom zależności liniowej między zmiennymi losowymi X i Y.
- *Współczynnika korelacji liniowej* definiuje się następująco:

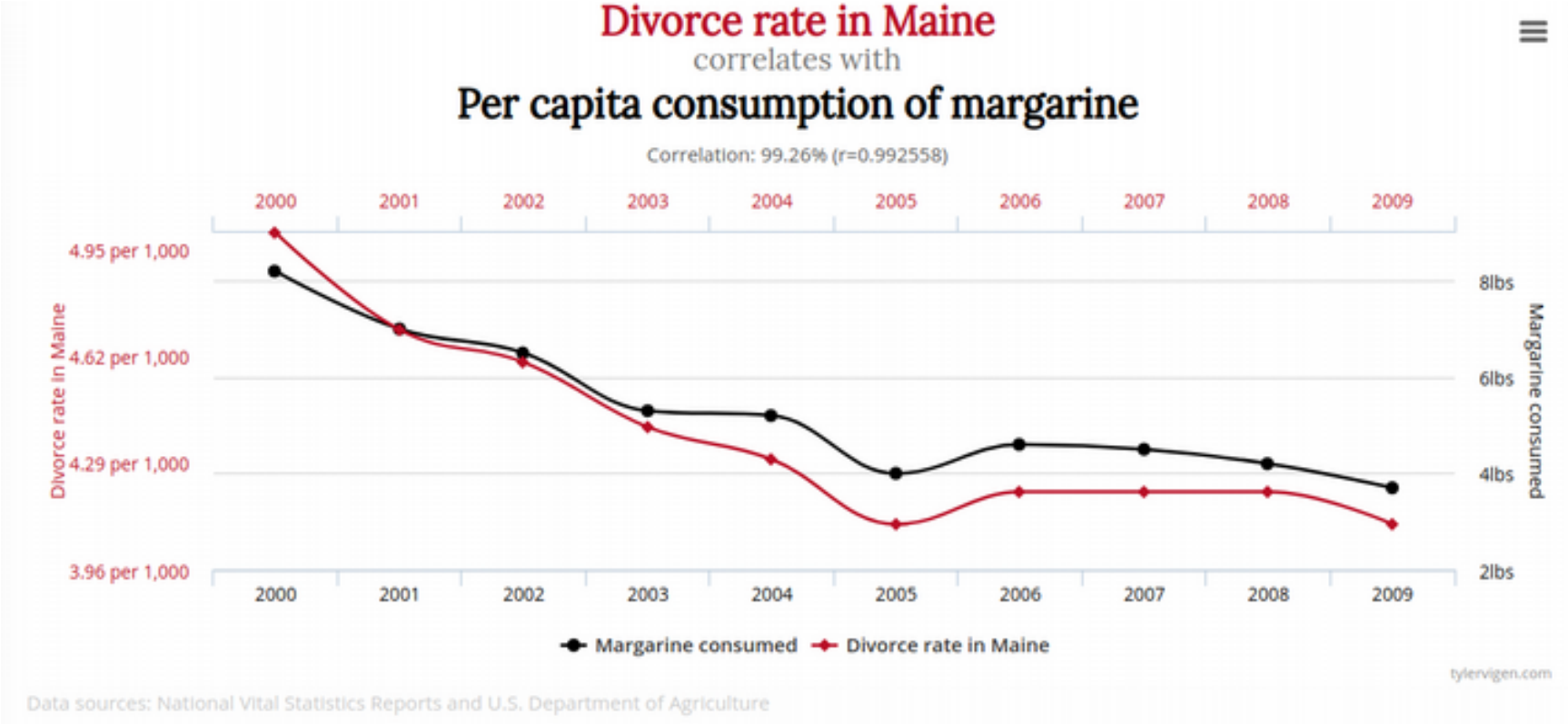
$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n}(x_i \cdot y_i) - \sum_{i=1}^{n}(x) \cdot \sum_{i=1}^{n}(y)}{(\sqrt{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

- Współczynnik korelacji liniowej dwóch zmiennych jest ilorazem *kowariancji* i iloczynu *odchyleń standardowych* tych zmiennych:

$$r_{xy} = \frac{cov(X,Y)}{\delta_X \cdot \delta_Y}$$

Korelacja - przykłady dziwnych korelacji

źródło: http://www.tylervigen.com/spurious-correlations



Analiza korelacji

- Analiza korelacji w statystyce polega na zbadaniu czy dwie zmienne są ze sobą istotnie statystycznie powiązane.
- W analizie korelacji nie bada się związku przyczynowo-skutkowego a po prostu związek/współwystępowanie dwóch zmiennych.
- Gdy badamy czy dwie zmienne są skorelowane ze sobą to nie wiemy, która zmienna wpływa na którą.

Dane

- **Dane** (zbiór danych) to zestaw wielu przykładów,
- Czasami przykłady określane są jako *punkty danych*,
- **Przykład** jest zbiorem *cech*, które zostały pomierzone na podstawie jakiegoś obiektu lub zdarzenia, a które system ma przetwarzać,
- Przykład można przedstawić jako wektor *cech*, *atrybutów*.

Dane

- W przypadku danych zakłada się że są one w jakiś sposób powiązane ze sobą (nie są chaotyczne),
- W wielu przypadkach mechanizm ten jest nieznany,
- W warunkach niepewności można przyjąć założenie, że mechanizm ten jest jakimś rozkładem prawdopodobieństwa.

Wiedza

- W przypadku wiedzy, możemy z pewnym przybliżeniem, twierdzić, że znane są nam jakieś "meta-związki", mechanizmy, które powodują powstanie danych

Rodzaje modeli

- *modele parametryczne* - gdy ilość parametrów określających model jest skończona
  np. możemy traktować parametry równania opisujące w/w krzywą jako model,
- *modele nieparametryczne* - gdy ilość parametrów opisujących model jest nieskończona
  np. możemy traktować krzywą (nieskończony zbiór punktów) jako model.
- W rzeczywistych problemach modele są tworzone, identyfikowane i używane w warunkach niepewności.

Modele parametryczne

- Parametryczne rozkłady prawdopodobieństwa. Najistotniejszym rozkładem jest rozkład normalny $N(\delta, m)$ opisywalny przez parę parametrów - *wariancję* δ i *średnią* m.
- Parametryczne modele regresyjne liniowe lub nieliniowe. Sposób w jaki zmienne zależą od siebie modelowany jest zazwyczaj w sposób jawny, strukturalizując ten wpływ.
- W tworzeniu modelu wybiera się arbitralnie sposób wpływu jednej zmiennej na drugą. Wykorzystuje się meta-wiedzę (nie zawartą w modelu).

Estymacja parametrów Θ modelu

- Estymator $\hat{\Theta}$ parametru Θ to funkcja próby $\hat{\Theta} = \{x_1, x_2, \ldots x_n\}$ która szacuje nieznaną wartość parametru rozkładu Θ.
- Estymator $\hat{\Theta}$ jako funkcja próby jest zmienną losową, ma więc swój rozkład, wartość oczekiwaną, wariancję, itp.
- Parametr Θ nie jest zmienną losową!
- Estymator $\hat{\Theta}$ parametru Θ jest zgodny, jeśli: $\hat{\Theta} \rightarrow_{n \rightarrow \infty} \Theta$

Modele nieparametryczne

- Histogram to jeden z graficznych sposobów przedstawienia rozkładu empirycznego cechy.
- Rozkłady posteriori mogą być reprezentowane poprzez wielokrotny pomiar tej samej wielkości.

Uczenie maszynowe

- Odpowiednie algorytmy mają pozwolić oprogramowaniu na zautomatyzowanie procesu pozyskiwania i analizy danych do ulepszania i rozwoju własnego systemu.
- Uczenie się może być rozpatrywane jako konkretyzacja algorytmu czyli dobór parametrów, nazywanych wiedzą lub umiejętnością. Służy do tego wiele typów metod pozyskiwania wiedzy oraz sposobów reprezentowania wiedzy.

Metody uczenia maszynowego

- uczenie z nadzorem,
- uczenie bez nadzoru,
- podział z nadzorem lub nie nie zawsze może być oczywisty.

Systemy uczące się

Mitchell (1997)

- ... program komputerowy uczy się na podstawie doświadczenia E względem pewnej klasy zadań T z wydajnością mierzoną P, jeśli działanie względem zadań T mierzone na podstawie P poprawia doświadczenie E..."
-

Miara wydajności modelu

- Zwykle miara wydajności jest powiązana z zadaniem wykonywanym przez system,
- np. dla zadań jak klasyfikacja jako wydajność modelu można przyjąć dokładność modelu (klasyfikacji) - procent przykładów dla których model działa dokładnie.
- współczynnik błędów - procent przykładów dla których wyniki modelu są niepoprawne

Zbiór testowy

- Zwykle jesteśmy zainteresowani, aby model uczył się danych, lecz w poprawny sposób uchwycił mechanizm generujący dane,
- Dlatego wyznaczamy miary wydajności za pomocą **zbioru testowego** danych - inny od danych używanych podczas nauki,
- Dane testowe jak i uczące powinny być "generowane"przez ten sam mechanizm.

Doświadczenie

- Algorytmy systemów uczących z grubsza można podzielić na **nadzorowane** i **nienadzorowane**,
- sposoby uczenia się:
  - **Nienadzorowane algorytmy uczące się -** poznają zbiór danych zawierających wiele cech, a następnie uczą się użytecznych właściwości dotyczących struktury tego zbioru danych,
  - **Nadzorowane algorytmy uczące się** poznają zbiór danych zawierających cechy, ale każdy przykład jest powiązany z *etykietą* czyli *celem*.

Dane szkoleniowe i testowe

Dane dzieli się na dane dwa rozłączne zbiory.

- **dane szkoleniowe -** dane w oparciu o które uczony jest model,
- **dane testowe -** dane służące do weryfikacji modelu,
- *Dane szkoleniowe* i *testowe* mają identyczny rozkład, wyprowadzony z tego samego rozkładu prawdopodobieństwa - **rozkładu generowania danych**,
- Przykłady w każdym zbiorze danych są od siebie *niezależne*.

Błąd szkoleniowy i uogólnienia

- **błąd szkoleniowy -** błąd obliczany dla danych szkoleniowych (np. suma błędów kwadratowych dla wzorców),
- **błąd uogólnienia (błąd testu) -** oczekiwana wartość błędu dla nowych danych wejściowych,
- Oczekiwany błąd szkoleniowy modelu powinien być równy oczekiwanemu błędowi testowemu tego modelu.

Nadmierne dopasowanie i niedopasowanie modelu

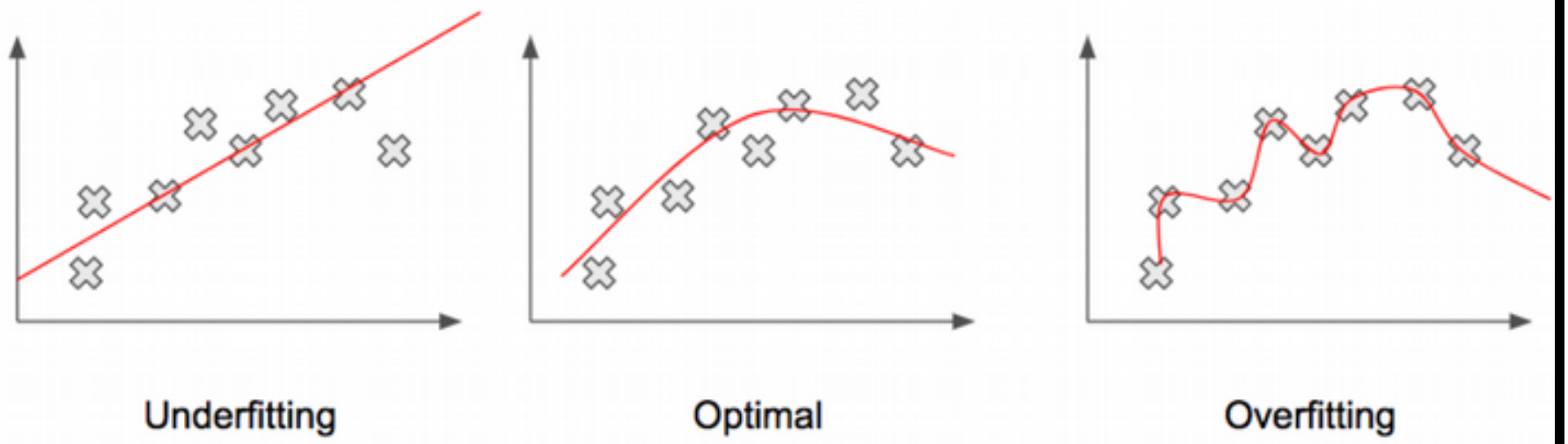
To czy dany algorytm będzie działać prawidłowo zależeć będzie od:

- Zdolności do sprawienia, aby błąd szkoleniowy był mały,
- Zdolności do sprawienia, aby różnica pomiędzy błędem szkolenia i błędem testu była mała.

Te dwa czynniki odpowiadają dwóm podstawowym wyzwaniom systemów uczących się:

- **nadmiernemu dopasowaniu**
- **nadmiernemu niedopasowaniu**

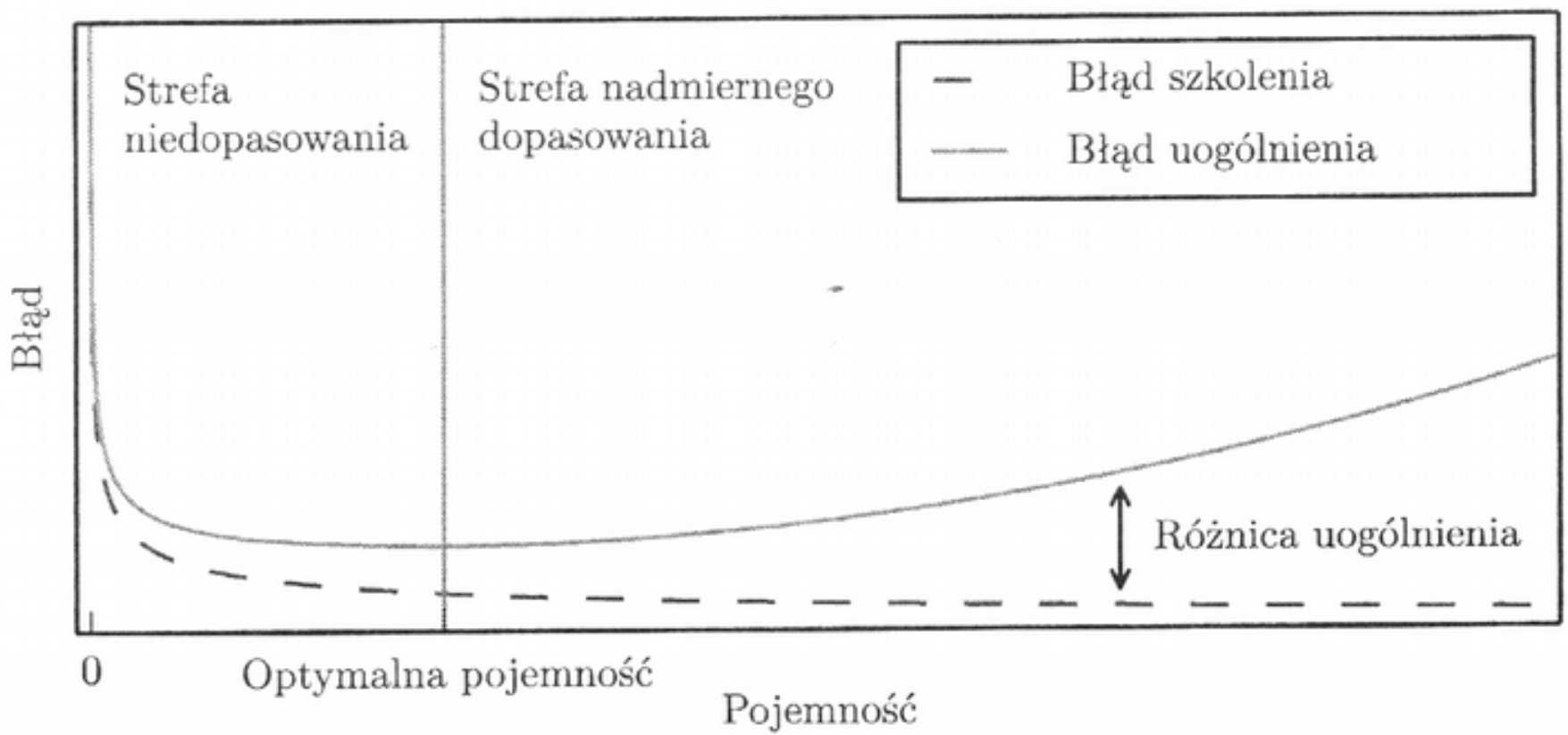
Nadmierne dopasowanie (overfitting) i niedopasowanie (underfitting)



Pojemność modelu

- **Pojemność modelu** - to jego zdolność do dopasowania do wielu różnych funkcji,
- Możemy kontrolować nadmierne dopasowanie modelu lub jego niedopasowanie zmieniając jego *pojemność*,
- Modele o niskiej pojemności mogą mieć kłopot z dopasowaniem zbioru szkoleniowego,
- Przykład: Wymiar Vapnika-Chervonenkisa (wymiar CV) dla binarnego klasyfikatora - największa możliwa wartość m, dla której istnieje szkoleniowy zbiór m różnych punktów x możliwych do arbitralnego oznaczenia.

Dopasowanie a pojemność modelu



Regularyzacja

- regularyzacja to jedna z technik, która może być użyta do kontrolowania zjawiska nadmiernego dopasowania,
- Polega ona na dodaniu “kary” do funkcji błędu, w celu ograniczenia dużych wartości błędu uogólnienia.

Zasada "Nie ma darmowych obiadów"

- **Zasada "Nie ma darmowych obiadów"** oznacza, że uśredniony względem wszystkich możliwych rozkładów generujących dane każdy algorytm klasyfikacyjny ma taki sam współczynnik błędów jak przy klasyfikacji punktów wcześniej nieobserwowanych (Wolpert 1996)
- Żaden algorytm dla systemów uczących nie jest uniwersalny,
- Algorytmy mogą działać dobrze w jednych dziedzinach, a winnych nie.

Sprawdzian krzyżowy

- Procedura ta opiera się na idei powtarzania obliczeń szkoleniowych i testowych na różnych losowo wybranych podzbiorach lub przedziałach oryginalnego zbioru danych,
- Najpopularniejszym z nich jest k-krokowy sprawdzian krzyżowy, w którym udział zbioru danych jest tworzony przez podzielenie go na k nienakładających się podzbiorów.