

Zaawansowane uczenie maszynowe: *podstawowe algorytmy uczenia się*

(rozszerzony przegląd)

Paweł Cichosz

Semestr 20Z

Przedstawiony tutaj przegląd podstawowych miar jakości modeli ma służyć jako pomocniczy materiał uzupełniający skrócony przegląd prezentowany na wykładzie 3 z przedmiotu *Zaawansowane uczenie maszynowe*. Zawiera on zwięzłe podsumowanie najważniejszych treści w punktach oraz dodatkowy szerszy komentarz w ramkach.

1 Miary jakości klasyfikacji

Błąd klasyfikacji na zbiorze: częstość pomyłek:

$$e_{S,c}(h) = \frac{|\{x \in S \mid h(x) \neq c(x)\}|}{|S|}$$

Podstawową miarą jakości klasyfikacji jest błąd klasyfikacji na zbiorze, czyli częstość, z jaką predykcje modelu różnią się od prawdziwych klas. Zazwyczaj występuje jednak potrzeba pogłębionej oceny jakości klasyfikacji, która nie ogranicza się do wyznaczenia częstości pomyłek (błędu), lecz bierze pod uwagę ich rozkład. Może nie być obojętne z punktu widzenia użyteczności modelu, jak często zdarzają się pomyłki dla poszczególnych klas. Zdarza się tak zwłaszcza w zadaniach o charakterze diagnostycznym (takich jak wykrywanie awarii, nadużyć, chorób itp.).

1.1 Macierz pomyłek

Macierz pomyłek na zbiorze S : $CM_{S,c}(h)[d_1, d_2] = |S_{h=d_1, c=d_2}|$

Macierz pomyłek (*confusion matrix*) reprezentuje rozkład pomyłek modelu na pewnym zbiorze przykładów. Jej wiersze i kolumny odpowiadają klasom, a w komórce na przecięciu wiersza odpowiadającego klasie d_1 i kolumny odpowiadającej klasie d_2 znajduje liczba pomyłek polegających na predykcji klasy d_1 dla przykładu, którego prawdziwą klasą jest d_2 .

Na podstawie macierzy pomyłek można oczywiście łatwo wyznaczyć błąd – jako iloraz sumy elementów poza główną przekątną do sumy wszystkich elementów. Jednak istotne korzyści przynosi dopiero uwzględnienie w ocenie jakości modelu niejednakowej częstości, z jaką model „myli się” dla różnych klas.

Przypadek dwuklasowy:

	c	
h	0	1
0	TN	FN
1	FP	TP

Miary jakości oparte na macierzy pomyłek definiuje się zwykle dla klasyfikacji binarnej (dwuklasowej). Jest to wprawdzie przypadek szczególny, ale odpowiadający znacznej części praktycznych zadań klasyfikacji o charakterze diagnostycznym, w których potrzeba analizy rozkładu pomyłek występuje najczęściej. Zobaczymy również (już na kolejnym wykładzie), w jaki sposób miary jakości, które zostaną dalej przedstawione i które dotyczą klasyfikacji dwuklasowej, mogą zostać wykorzystane także przy ocenie modeli klasyfikacji wieloklasowej.

Wygodnie jest formułować definicje miar jakości klasyfikacji opartych na macierzy pomyłek z wykorzystaniem umownych, lecz przyjętych powszechnie nazw jej elementów: TP (*true positive*, prawdziwe pozytywne), TN (*true negative*, prawdziwe negatywne), FP (*false positive*, fałszywe pozytywne) i FN (*false negative*, fałszywe negatywne). Konwencja polega na tym, że *drugi* człon nazwy (P/N, pozytywne/negatywne) oznacza odpowiednio klasę 1 albo 0 przewidywaną przez model, a *pierwszy* człon nazwy (T/F, true/false, prawdziwe/fałszywe) – mówi o tym, czy ta przewidywana klasa jest zgodna z klasą prawdziwą. TP oznacza więc liczbę przykładów, które zostały przez model zaklasyfikowane do klasy 1 i naprawdę należą do tej klasy, FP – liczbę przykładów, które zostały przez model zaklasyfikowane do klasy 1, lecz naprawdę należą do klasy 0, itd.

Jest kwestią konwencji, którą z dwóch klas w praktycznym zadaniu klasyfikacji uznaje się za pozytywną, a którą za negatywną. Ponieważ, jak zobaczymy, często stosowane miary jakości w pewnym sensie wyróżniają klasę pozytywną, naturalne jest uznanie za pozytywną klasy „bardziej interesującej” w rozważanym zastosowaniu, czyli takiej, do której wykrywania ma służyć model.

1.2 Miary jakości oparte na macierzy pomyłek

błąd:

$$\frac{FP + FN}{TP + TN + FP + FN}$$

dokładność:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Dla porządku możemy na podstawie macierzy pomyłek 2×2 zdefiniować na początek błąd oraz jego dopełnienie do 1, czyli dokładność. Bardziej interesujące są jednak inne miary, odzwierciedlające różnice między liczbą pomyłek różnego rodzaju.

współczynnik prawdziwych pozytywnych (true positive rate):

$$\frac{TP}{TP + FN}$$

współczynnik fałszywych pozytywnych (false positive rate):

$$\frac{FP}{TN + FP}$$

Współczynnik prawdziwych pozytywnych albo *TP rate* jest zdefiniowany jako stosunek liczby przykładów pozytywnych prawidłowo klasyfikowanych jako pozytywne do liczby wszystkich przykładów pozytywnych. Mówi on o skuteczności, z jaką model wykrywa przykłady pozytywne, co jest zazwyczaj głównym przedmiotem zainteresowania w zastosowaniach diagnostycznych. Jego maksymalizacja może być jednak w trywialny sposób osiągnięta przez bezużyteczny model, który zawsze przewiduje klasę 1, a więc nie może on być jedyną podstawą oceny jakości.

Wskaźnikiem komplementarnym do *TP rate* jest współczynnik fałszywych pozytywnych (*FP rate*), który jest obliczany jako częstość niepoprawnej predykcji klasy 1 dla przykładów w rzeczywistości negatywnych. Ten wskaźnik powinien być minimalizowany, a można to w trywialny sposób osiągnąć przez predykcję zawsze klasy 0.

Para wskaźników *TP rate* i *FP rate* używanych łącznie jednak już bardzo dobrze charakteryzuje jakość predykcji. Ponieważ występuje między nimi oczywista wymiana (pożądane zwiększanie *TP rate* zwykle pociąga za sobą niepożądane zwiększanie *FP rate*, pożądana redukcja *FP rate* zwykle pociąga za sobą niepożądaną redukcję *TP rate*), konieczne jest szukanie pewnego poziomu równowagi. Proste strategie mogą polegać na nałożeniu na jeden ze wskaźników ograniczenia (np. warunku, aby *FP rate* nie przekraczało pewnego poziomu) i przy tym ograniczeniu optymalizowaniu drugiego (np. maksymalizacji *TP rate*). W dalszym ciągu wykładu zobaczymy, że szczególnie wygodne jest poszukiwanie takiego kompromisu z wykorzystaniem pewnej metody wizualizacji.

odzysk (recall): = współczynnik prawdziwych pozytywnych

precyzja (precision):

$$\frac{TP}{TP + FP}$$

miara F : średnia harmoniczna precyzji i odzysku:

$$F = \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

Inną parą komplementarnych wskaźników, które rozpatrywane łącznie charakteryzują jakość predykcji binarnych, jest para odzysk (*recall*) i precyzja (*precision*). Pierwszy z nich jest synonimem TP rate. Drugi jest obliczany jako stosunek liczby przykładów, dla których model przewiduje klasę 1 i które faktycznie należą do tej klasy do liczby wszystkich przykładów, dla których model przewiduje klasę 1. Jest to więc precyzja, z jaką model „alarmuje” – wartość mówiąca o tym, jak duża część „alarmów” (predykcji 1) okazuje się trafna.

W tym przypadku oba wskaźniki powinny być maksymalizowane, co niekiedy może być wygodniejsze. Umożliwia w szczególności łatwe zagregowanie ich do jednego wskaźnika, zdefiniowanego jako ich średnia. Przyjęte jest jednak stosowanie w tym celu średniej harmonicznej, na którą większy wpływ ma mniejsza z dwóch uśrednianych wartości. Średnią harmoniczną odzysku i precyzji nazywa się miarą F (czasem $F1$ lub F_1) i niekiedy wykorzystuje jako wygodne kryterium porównywania jakości modeli, chociaż oczywiście kompresja macierzy pomyłek do jednego wskaźnika oznacza istotną utratę informacji i modele o jednakowych wartościach miary F mogą mieć wciąż istotnie różny rozkład pomyłek.

czułość (sensitivity): = współczynnik prawdziwych pozytywnych

specyficzność (specificity): = $1 -$ współczynnik fałszywych pozytywnych

Warto wspomnieć również o parze wskaźników czułość i specyficzność, z których pierwszy jest kolejnym synonimem TP rate, a drugi – dopełnieniem do 1 FP rate. Jest to różnica czysto techniczna i wskaźniki te nie wnoszą nic ponad to, czego dostarczają TP rate i FP rate. Wspominamy o nich, gdyż można jednak dość często spotkać je w literaturze lub w oprogramowaniu do oceny jakości modeli.

1.3 Analiza ROC

Graficzna metoda oceny jakości klasyfikacji: wizualizacja punktów pracy modeli klasyfikacji za pomocą punktów i krzywych w układzie współrzędnych (TP rate [y], FP rate [x]).

Analiza ROC (*receiver operating characteristic* – metoda wywodzi się z badań dotyczących detekcji obiektów na podstawie sygnałów radarowych prowadzonych w czasie II wojny światowej) jest metodą ułatwiającą śledzenie wymiany i poszukiwanie kompromisu między TP rate i FP rate, a także porównywanie modeli. Wskaźniki te odpowiadają w tym podejściu osiom kartezjańskiego układu współrzędnych (TP rate – oś y , FP rate – oś x).

Predykcja klas: pojedynczy punkt.

Jeśli ocenie na pewnym zbiorze danych podlega model dokonujący predykcji klas, to uzyskane wartości TP rate i FP rate są reprezentowane przez jeden punkt na płaszczyźnie ROC.

Predykcja probabilistyczna: krzywa (łamana) łącząca punkty odpowiadające różnym progom odcięcia $P(1|x)$.

W przypadku, gdy oceniany jest model dokonujący predykcji probabilistycznych, stosując różne progowe wartości prawdopodobieństwa $P(1|x)$, powyżej których przykładowi przypisywana jest klasa 1, otrzymuje się różne etykietowania przykładów i odpowiednio różne wartości wskaźników TP rate i FP rate. Każda osiągalna przy przesuwaniu progu odcięcia między 0 a 1 para wartości TP rate, FP rate reprezentuje pewien możliwy *punkt pracy* modelu klasyfikacji. Zaznaczając wszystkie osiągalne punkty pracy i łącząc je odcinkami otrzymujemy tzw. krzywą ROC (w rzeczywistości łamaną). Stanowi ona wizualną reprezentację jakości klasyfikacji probabilistycznego modelu klasyfikacji.

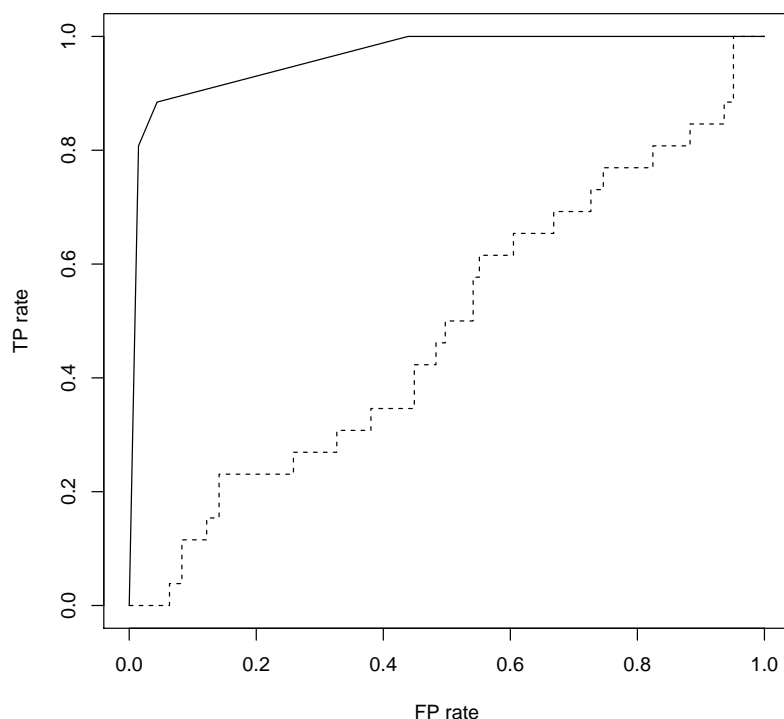
Ścisłe rzecz biorąc, analizę ROC można zresztą prowadzić dla dowolnych modeli generujących predykcje liczbowe, które służą do przypisania binarnych etykiet klas na podstawie porównania z pewnym progiem – nie jest konieczne, aby te liczbowe predykcje mieściły się w przedziale $[0, 1]$ i miały interpretację probabilistyczną, chociaż najczęściej tak się właśnie przyjmuje.

Liczba punktów pracy: powiększona o 1 liczba różnych wartości $P(1|x)$.

Na każdym wykresie ROC są przynajmniej dwa skrajne punktu $(0, 0)$ i $(1, 1)$. Wszystkich możliwych punktów pracy na krzywej ROC jest tyle, ile wynosi liczba różnych wartości $P(1|x)$ powiększona o 1. Model (oczywiście bezużyteczny), który wszystkim przykładom przypisuje jednakową wartość tego prawdopodobieństwa, ma tylko dwa punkty pracy: $(0, 0)$ i $(1, 1)$, a jego krzywa ROC przebiega przez przekątną kwadratu jednostkowego.

Podobny przebieg ma krzywa dla modelu, który generuje losowe predykcje prawdopodobieństw, pozbawione jakiegokolwiek związku z prawdziwymi klasami. Wówczas każdemu wzrost TP rate towarzyszy mniej więcej taki sam wzrost FP rate (z dokładnością do losowych odchyłek) i krzywa ROC oscyluje wokół przekątnej. Krzywe ROC dla modeli o ponadlosowej jakości predykcji są położone powyżej głównej przekątnej – ich punkty pracy charakteryzują się wyższą wartością TP rate niż FP rate. Z kolei krzywe ROC znajdujące się zdecydowanie poniżej głównej przekątnej zwykle wskazują na techniczny błąd etykietowania (zamianę etykiet klas pozytywnej i negatywnej) przy sporządzaniu wykresu.

Rys. 1 przedstawia przykładową krzywą ROC dla modelu o wysokiej jakości predykcji oraz dla modelu generującego losowe predykcje.



Rysunek 1: Przykładowe krzywa ROC dla modelu o wysokiej jakości predykcji i dla modelu generującego losowe predykcje.

Zagregowana ocena predykcji probabilistycznych: pole pod krzywą ROC (AUC) – interpretowane jako prawdopodobieństwo, że losowo wybrany przykład klasy 1 uzyska większe $P(1|x)$ niż losowo wybrany przykład klasy 0.

Mówiąc nieformalnie, korzystne są krzywe położone „jak najwyżej” i szybko rosnące – wówczas uzyskiwany jest znaczący przyrost TP rate przy dość niewielkim wzroście FP rate. Często stosowaną zagregowaną miarą jakości wszystkich osiągalnych punktów pracy jest pole pod krzywą ROC, tzn. AUC (*area under curve*).

Pole pod krzywą ROC ma dość intuicyjną interpretację: reprezentuje ono prawdopodobieństwo tego, że losowy wybrany przykład klasy 1 uzyska według modelu wyższą wartość $P(1|x)$ niż losowo wybrany przykład klasy 0. W typowych zastosowaniach uznaje się za „dobre” lub „bardzo dobre” modele o wartościach AUC powyżej 0.85, za „dość dobre” modele o wartościach AUC powyżej 0.7, są to jednak tylko orientacyjne wskazania. W stosunkowo łatwych zadaniach predykcyjnych i przy wysokim poziomie wymagań jakości predykcji nawet poziom 0.9 może nie być zadowalający, a w przypadku predykcji dotyczącej zjawisk bardzo trudno przewidywalnych nawet ledwie ponadlosowy poziom AUC ok. 0.6 może być czasem uważany za sukces.

Warto zauważyć, że nie zawsze model o najwyższej wartości AUC osiąga najbardziej preferowany punkt pracy – model o nieco niższej wartości AUC może w początkowej części krzywej mieć przewagę. Niekiedy stosuje się do oceny modeli wartość AUC ograniczoną do zakresu „potencjalnie użytecznych” punktów pracy, dla których FP rate nie przekracza 0.5.

Sporządzanie wykresu: sortowanie według $P(1|x)$, wyznaczanie zmiany TP i FP przy każdej zmianie progu odcięcia.

„Ręczne” sporządzanie wykresu ROC (oczywiście dla miniatury zbiorów danych) jest dość proste. Wystarczy zbiór danych uporządkować według wartości $P(1|x)$ (np. niemalejąco) oraz rozważać kolejno możliwe progi odcięcia w zakresie od 0 do 1 (w dowolnym kierunku). W celu identyfikacji wszystkich możliwych pośrednich punktów pracy dla każdej pary sąsiednich różnych wartości $P(1|x)$ należy rozważyć jeden próg odcięcia położony między nimi.

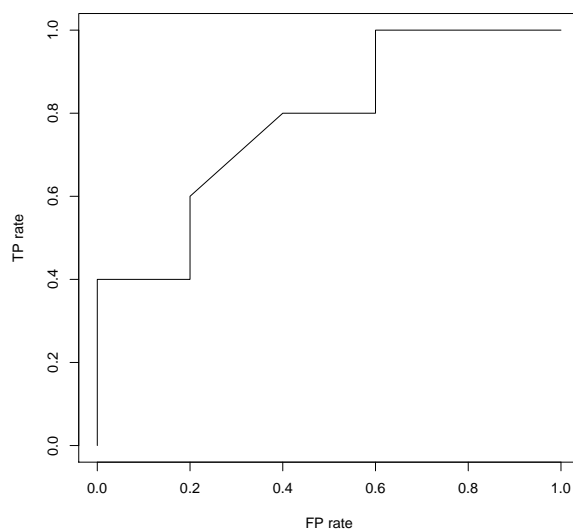
Jeśli zaczniemy od minimalnego progu odcięcia, to odpowiada mu punkt (1,1) – dla wszystkich przykładów przewidywane prawdopodobieństwo klasy 1 przekracza próg, a więc wszystkie otrzymują przewidywaną klasę 1, co powoduje uzyskanie maksymalnych wartości TP rate i FP rate równych 1. Odpowiednio TP (liczba wszystkich przykładów pozytywnych poprawnie klasyfikowanych jako pozytywne) jest równe liczbie wszystkich przykładów klasy 1, a FP (liczba wszystkich negatywnych niepoprawnie klasyfikowanych jako pozytywne) jest równa liczbie wszystkich przykładów klasy 0.

Następnie należy przesunąć się z progiem odcięcia do kolejnej możliwej wartości, co powoduje zmianę przewidywanej klasy (z 1 na 0) dla jednego lub większej liczby przykładów. Te z nich, które mają prawdziwą klasę 1, pomniejszają TP, zaś te o prawdziwej klasie 0 – pomniejszają FP. Skorygowane TP i FP umożliwiają wyznaczenie wartości

wskaźników TP rate i FP rate dla uzyskanego nowego punktu pracy. Postępując dalej w ten sam sposób docieramy w końcu do maksymalnego progu odcięcia, któremu odpowiada punkt $(0, 0)$ – dla wszystkich przykładów przewidywane prawdopodobieństwo klasy 1 nie przekracza progu, a więc wszystkie otrzymują przewidywaną klasę 0, co powoduje uzyskanie minimalnych wartości TP rate i FP rate równych 0.

Rys. 2 przedstawia przykładową krzywą ROC dla modelu z ośmioma punktami pracy, uzyskaną na podstawie wartości $P(1|x)$ i klas przedstawionych w zamieszczonej obok tabeli.

$P(1 x)$	$c(x)$
0.1	0
0.1	0
0.3	1
0.4	0
0.5	0
0.5	1
0.6	1
0.7	0
0.9	1
0.9	1



Rysunek 2: Przykładowa krzywa ROC dla modelu z ośmioma punktami pracy.

2 Miary jakości regresji

Model regresji ma za zadanie przybliżać wartości funkcji docelowej. Jakość tego przybliżenia nie będzie już oceniana przez zliczanie lub estymowanie prawdopodobieństwa pomyłek, lecz raczej na podstawie różnic między prawdziwymi i przewidywanymi wartościami, które nazywane są *residuami* bądź *resztami* modelu.

Podane miary jakości regresji mogą być wyznaczane na dowolnym zbiorze danych, na którym znane są prawdziwe wartości funkcji docelowej. Oczywiście jednak, gdy interesuje nas ocena oczekiwanej jakości modelu na nowych danych, powinien to być zbiór rozłączny ze zbiorem trenującym. Taka ocena jest prowadzona przez zastosowanie

takich samych procedur oceny, jakie omawialiśmy przy okazji modeli klasyfikacji (np. *holdout* lub *k-CV*).

Residuum (reszta): różnica $f(x) - h(x)$.

Błąd bezwzględny (MAE, *mean absolute error*):

$$\text{mae}_{S,f}(h) = \frac{1}{|S|} \sum_{x \in S} |f(x) - h(x)|$$

Błąd średniokwadratowy (MSE, *mean square error*):

$$\text{mse}_{S,f}(h) = \frac{1}{|S|} \sum_{x \in S} (f(x) - h(x))^2$$

Podstawowe miary jakości regresji oparte na agregowaniu residuów to błąd bezwzględny (MAE) i błąd średniokwadratowy (MSE). Pierwszy jest średnią wartości bezwzględnych residuów, a drugi – średnią kwadratów residuów uzyskanych na zbiorze przykładów, na którym przeprowadzana jest ocena. Błąd średniokwadratowy w większym stopniu „karze” duże różnice, a w dodatku ma lepsze właściwości analityczne (różniczkowalność, która jest wykorzystywana przez niektóre algorytmy), co powoduje, że jest częściej stosowany.

Pierwiastek błędu średniokwadratowego (RMSE, *root mean square error*):

$$\text{rmse}_{S,f}(h) = \sqrt{\text{mse}_{S,f}(h)}$$

W celu łatwiejszej interpretacji błędu średniokwadratowego czasem podaje się jego pierwiastek (RMSE). Jeśli predykcja dotyczy wielkości wyrażonych w pewnych jednostkach (fizycznych, monetarnych itp.), to MSE wyraża błąd w kwadratach tych jednostek, a RMSE sprowadza go ponownie do pierwotnych jednostek. Ma to jednak oczywiście znaczenie tylko do celów prezentacji oceny jakości modeli, a nie wpływa na to, który z porównywanych modeli będzie uznany za lepszy.

Błąd względny (RAE, *relative absolute error*):

$$\text{rae}_{S,f}(h) = \frac{\sum_{x \in S} |f(x) - h(x)|}{\sum_{x \in S} |f(x) - m_S(f)|} = \frac{\text{mae}_{S,f}(h)}{\frac{1}{|S|} \sum_{x \in S} |f(x) - m_S(f)|}$$

Współczynnik determinacji:

$$R^2_{S,f}(h) = 1 - \frac{\sum_{x \in S} (f(x) - h(x))^2}{\sum_{x \in S} (f(x) - m_S(f))^2} = 1 - \frac{|S| \text{mse}_{S,f}(h)}{(|S| - 1) s_S^2(f)}$$

Łatwiejszej interpretacji służy błąd względny (RAE), który jest zdefiniowany jako iloraz średniego błędu bezwzględnego i średniej bezwzględnej różnicy między wartościami atrybutu docelowego a ich średnią. W ten sposób jakość predykcji za pomocą modelu jest normalizowana za pomocą jakości predykcji za pomocą średniej. Wartości bliskie 1 lub powyżej 1 wskazują na bezużyteczny model.

W podobny sposób błąd średniokwadratowy może być normalizowany za pomocą wariancji, tzn. suma kwadratów różnic między wartościami prawdziwymi i przewidywanymi może być podzielona między sumę kwadratów różnic między wartościami prawdziwymi a ich średnią. Przyjęte jest stosowanie dopełnienia do 1 tego ilorazu, nazywane współczynnikiem determinacji. Mimo zwyczajowego oznaczenia R^2 wartość tego współczynnika może być ujemna jeśli model przewiduje gorzej niż średnia. Wartości bliskie 0 lub ujemne wskazują na bezużyteczne modele, a wartości bliskie 1 na niemal doskonałą predykcję. O współczynniku determinacji mówi się, że wyraża "część wariancji atrybutu docelowego objaśnianą przez model".

Średnia strata:

$$\frac{1}{|S|} \sum_{x \in S} \mathcal{L}(f(x), h(x))$$

gdzie \mathcal{L} – funkcja straty, np.:

$$\begin{aligned} \mathcal{L}(f(x), h(x)) &= (f(x) - h(x))^2 && \text{– strata kwadratowa} \\ \mathcal{L}(f(x), h(x)) &= |f(x) - h(x)| && \text{– strata bezwzględna} \\ \mathcal{L}(f(x), h(x)) &= \mathbb{I}_{h(x) \neq f(x)} && \text{– strata zero-jedynkowa} \end{aligned}$$

($\mathbb{I}_{\text{warunek}}$ oznacza indyktor równy 1 gdy warunek jest spełniony i 0 w przeciwnym przypadku)

Błąd średniokwadratowy i średni błąd bezwzględny są oparte na sumowaniu kwadratów bądź bezwzględnych wartości residuów. Można je traktować jako szczególne przypadki tzw. *straty*, która wyraża udział predykcji dla poszczególnych przykładów w ocenie jakości modelu. W ogólności funkcja straty dla prawdziwej wartości $f(x)$ i przewidywanej wartości $h(x)$ wyraża liczbowy koszt związany z tą predykcją przy tej prawdziwej wartości, czyli jest odpowiednikiem kosztów pomyłek rozważanych przez nas wcześniej dla modeli klasyfikacji. Szczególne przypadki funkcji straty to właśnie strata kwadratowa (kwadrat różnicy $f(x) - h(x)$) oraz strata bezwzględna (wartość bezwzględna różnicy

$f(x) - h(x)$). Błąd średniokwadratowy to średnia wartość straty kwadratowej, a błąd bezwzględny – średnia wartość straty bezwzględnej.

Zauważmy, że funkcję straty możemy również stosować w przypadku zadań klasyfikacji. W szczególności tzw. strata zerojedynkowa, która jest równa 1 gdy $f(x) \neq h(x)$ i 0 gdy $f(x) = h(x)$ odpowiada ocenie modelu klasyfikacji za pomocą błędu klasyfikacji (błąd klasyfikacji to średnia wartość straty zerojedynkowej).

Korelacja: współczynnik korelacji liniowej lub rangowej między wartościami f i h .

W roli miary jakości modelu regresji bywa czasem także stosowany współczynnik korelacji liniowej lub rangowej między wartościami prawdziwymi i przewidywanymi. Ma to uzasadnienie, kiedy o przydatności modelu do przewidywanego zastosowania nie decyduje wielkość różnic między f i h , lecz to, czy model trafnie rozróżnia wartości większe i mniejsze. Taka potrzeba występuje w zastosowaniach, w których model regresji ma np. rozróżniać między przykładami (np. pracownikami, sklepami, urządzeniami, instalacjami przemysłowymi itp.) „lepszymi” i „gorszymi” ze względu na pewną miarę jakości (np. wydajność pracy, wielkość produkcji, wielkość sprzedaży itp.).