Eksploracja zasobów internetowych

Wykład 5

Klasyfikowanie dokumentów WWW

Piotr Hońko
Wydział Informatyki
Politechnika Białostocka

Klasyfikowanie danych

- Znajdowanie sposobu odwzorowania zbioru obiektów w zbiór predefiniowanych klas, gdzie klasy są ustalone w sposób naturalny lub przez eksperta.
- Automatyczne kategoryzowanie nowych danych z wykorzystaniem sposobu odwzorowania obiektów w klas.

Klasyfikacja – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	*30*	*male*	*1800*	*no*
2	*33*	*female*	*2500*	*yes*
3	*30*	*male*	*2000*	*no*
4	*30*	*female*	*1800*	*yes*
5	*26*	*female*	*2500*	*yes*
6	*30*	*male*	*3000*	*yes*
7	*30*	*female*	*1600*	*no*

- Znaleźć odwzorowanie zbioru obiektów {*1*,*2*,..,*7*}, scharakteryzowanych przez atrybuty *Age*, *Gender*, *Income*, w zbiór klas określonych przez atrybut *GoodCustomer*.

Zadanie klasyfikacji – etapy

1. Budowa klasyfikatora.
2. Weryfikacja jakości klasyfikatora:
    1. wstępna – wykorzystanie danych treningowych;
    2. właściwa – wykorzystanie danych testowych.
3. Zastosowanie klasyfikatora do kategoryzacji nowych danych.

Klasyfikator

- Klasyfikator to mechanizm, który na podstawie wiedzy nabytej w procesie uczenia lub eksploracji danych potrafi określić dla dowolnego obiektu z bazy danych klasę, do której obiekt powinien być przypisany.

Weryfikacja jakości klasyfikatora

- Klasyfikator jest oceniany za pomocą miary, która na podstawie rzeczywistych i proponowanych przez klasyfikator klas obiektów jest w stanie określić jego jakość/przydatność.
- Podstawowe miary oceny jakości klasyfikatora:
  - dokładność – stosunek liczby obiektów poprawnie; sklasyfikowanych do wszystkich obiektów sklasyfikowanych
  - pokrycie – stosunek liczby obiektów poprawnie sklasyfikowanych do wszystkich obiektów;
  - czułość – stosunek liczby obiektów wybranej klasy, poprawnie sklasyfikowanych do wszystkich obiektów wybranej klasy;
  - precyzja – stosunek liczby obiektów wybranej klasy, poprawnie sklasyfikowanych do wszystkich obiektów przypisanych do danej klasy.

Metody oceny jakości klasyfikatora

- Ponowne podstawienie (ang. *resubstitution*)
  - Klasyfikator jest generowany i testowany na podstawie tych samych danych.
- Trenuj i testuj (ang. *train and test*)
  - Zbiór danych dzielony jest na dwie części – treningową i testową. Klasyfikator generowany jest na podstawie danych treningowych, a oceniany na danych testowych.

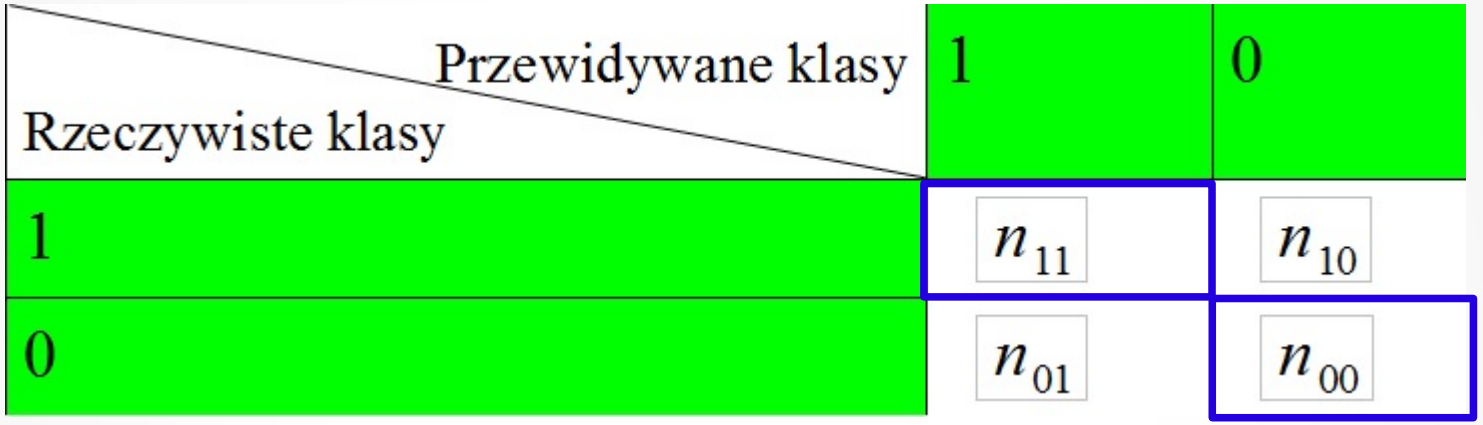
Metody oceny jakości klasyfikatora c.d.

- Krzyżowa walidacja (ang. *cross-validation*)
  - Podział zbioru danych na ustaloną z góry liczbę *k* równolicznych podzbiorów.
  - Przeprowadzenie *k* testów metodą trenuj i testuj (zbiór testowy – jeden z *k* podzbiorów; zbiór treningowy – suma pozostałych podzbiorów).
  - Uśrednienie wyniku oceny jakości klasyfikatora na podstawie *k* testów.

Rodzaje klasyfikacji

- Klasyfikacja binarna
  - Dotyczy pojedynczego pojęcia, np. pojęcie dobrego klienta;
  - Baza zwiera przykłady pozytywne i negatywne danego pojęcia, przez co ustalany jest podział na dwie klasy.
- Klasyfikacja wieloklasowa
  - Dotyczy wielu pojęć, np. trzy gatunki irysów.
  - Baza zawiera (zwykle) tylko przykłady pozytywne każdego z pojęć, przez co ustalany jest podział na tyle klas, ile jest różnych pojęć.

Klasyfikacja binarna



n_{ij} – liczba obiektów klasy *i* przypisanych do klasy *j*.

Klasyfikacja wieloklasowa

Rzeczywiste klasy \ Przewidywane klasy	c_1	c_2	…	c_k
c_1	$n_{c_1c_1}$	$n_{c_1c_2}$	…	$n_{c_1c_k}$
c_2	$n_{c_2c_1}$	$n_{c_2c_2}$	…	$n_{c_2c_k}$
…	…	…	…	…
c_k	$n_{c_kc_1}$	$n_{c_kc_2}$	…	$n_{c_kc_k}$

$n_{c_ic_j}$ – liczba obiektów klasy c_i przypisanych do klasy c_j .

Rodzaje klasyfikatorów

- Reguły decyzyjne
- Drzewa decyzyjne
- Sieci neuronowe
- Sieci bayesowskie
- Maszyna wektorów podpierających (ang. *Support Vector Machine*)
- Klasyfikatory oparte o uczenie z przykładów (ang. *Instance Based Learning*)
- …

Reguła klasyfikacyjne (decyzyjne)

- Reguła klasyfikacyjna jest wyrażaniem postaci:

$$(a_1, v_1) \wedge (a_2, v_2) \wedge \ldots \wedge (a_n, v_n) \rightarrow (d, v)$$

gdzie

- $a_i \, (1 \leq i \leq n)$ – atrybuty warunkowe
- $v_i \, (1 \leq i \leq n)$ – wartości atrybutów warunkowych,
- d – jest atrybutem decyzyjnym,
- v – wartość atrybutu decyzyjnego (klasa decyzyjna).

- Przy konstrukcji reguły zamiast wartości atrybutu może być rozpatrywany zbiór wartości.

Reguła klasyfikacyjne – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	*30*	*male*	*1800*	*no*
2	*33*	*female*	*2500*	*yes*
3	*30*	*male*	*2000*	*no*
4	*30*	*female*	*1800*	*yes*
5	*26*	*female*	*2500*	*yes*
6	*30*	*male*	*3000*	*yes*
7	*30*	*female*	*1600*	*no*

$(Income, [2500,3000]) \rightarrow (GoodCustomer, yes)$ (dokładność=1, pokrycie=3/4)

$(Gender, Female) \wedge (Income, [1800,3000]) \rightarrow (GoodCustomer, yes)$

(dokładność=1, pokrycie=3/4)

Drzewa klasyfikacyjne (decyzyjne)

- Drzewo klasyfikacyjne to skierowany, spójny graf acykliczny.
- Drzewo klasyfikacyjne składa się z:
  - węzłów zawierających testy przeprowadzone na atrybutach;
  - gałęzi zawierających wyniki testów;
  - liści zawierających klasy decyzyjne.

Drzewa klasyfikacyjne – przykład

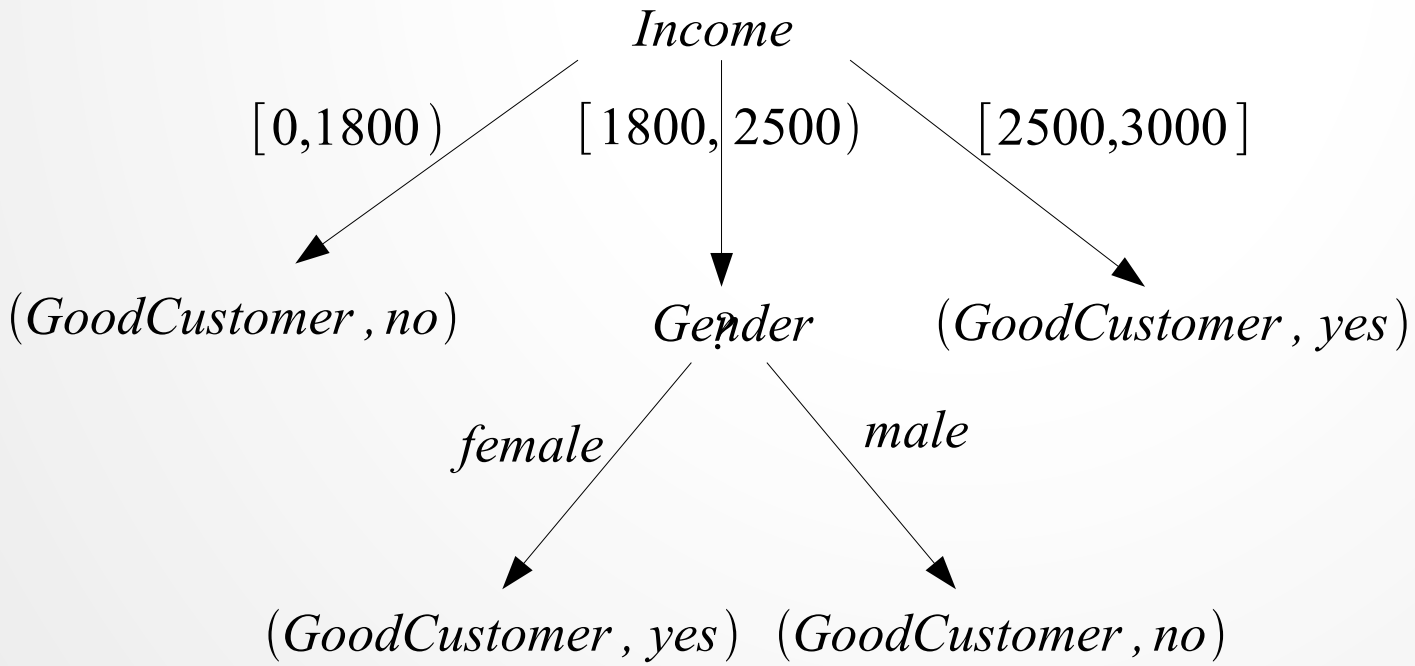
Customer

id	Age	Gender	Income	GoodCustomer
1	30	male	1800	no
2	33	female	2500	yes
3	30	male	2000	no
4	30	female	1800	yes
5	26	female	2500	yes
6	30	male	3000	yes
7	30	female	1600	no

Jedna z reguł utworzona na podstawie drzewa:

$$(Income, [1800,2500)) \wedge (Gender, female) \rightarrow (GoodCustomer, yes)$$

(dokładność=1, pokrycie=3/4)



Uczenie z przykładów

- Proces generowania wzorców jest pomijany lub ograniczony do zapamiętania (wybranych) obiektów treningowych.
- Nowe obiekty są klasyfikowane na podstawie ich powiązań z zapamiętanymi obiektami.
- Związek ten określany jest zwykle za pomocą miar odległości lub podobieństwa.

Ocena jakości uczenia z przykładów

- Ocena jakości klasyfikatora nie jest możliwa, do momentu przeprowadzenia klasyfikacji danych testowych.
- Próbka zapamiętanych obiektów treningowych może być poddana częściowej ocenie jakości, np. poprzez dokonanie oceny statystycznej.

Uczenie z przykładów – metoda *k* najbliższych sąsiadów

- Obiekt klasyfikowany jest na podstawie głosowań *k* najbliższych mu obiektów treningowych.
- Każdy z *k* obiektów oddaje głos na klasę, do której należy.
- Nowy obiekt jest przypisywany do klasy, która wygrywa w głosowaniu.

Metoda *k* najbliższych sąsiadów – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	30	male	1800	no
2	33	female	2500	yes
3	30	male	2000	no
4	30	female	1800	yes
5	26	female	2500	yes
6	30	male	3000	yes
7	30	female	1600	no

Miara odległości:

$$d(o_1,o_2)=\frac{|Age(o_1)-Age(o_2)|}{max(Age)-min(Age)}+1-(Gender(o_1)=Gender(o_2))+\frac{|Income(o_1)-Income(o_2)|}{max(Income)-min(Income)}$$

Odległość obiektu 8 od każdego obiektu bazy:

1	0,39	+
2	1,86	
3	0,25	+
4	1,39	
5	2,61	
6	0,96	+
7	1,54	

- Niech *k*=3.
- Dokonujemy klasyfikacji obiektu (*8*,*31*,*male*,*2000*,*?*)
- Wynik głosowania: *yes* – *no*: 1 – 2;
- Obiekt zostaje przypisany do klasy *no*.

Klasyfikacja dokumentów WWW

- Znajdowanie sposobu odwzorowania zbioru dokumentów w zbiór predefiniowanych klas.
- Klasy dokumentów są ustalane według dowolnego kryterium, na podstawie którego dokumenty mogą być zorganizowane w klasy (np. tematyka stron).

Ogólny schemat procesu klasyfikacji dokumentów

1. Gromadzenie i przetwarzanie danych dotyczących dokumentów.
2. Budowanie modelu odwzorowania dokumentów w klasy.
3. Testowanie i ocena modelu.
4. Klasyfikacja nowych dokumentów za pomocą otrzymanego modelu.

Gromadzenie i przetwarzanie danych dotyczących dokumentów

- Gromadzenie dokumentów w postaci tekstowej
- Oczyszczanie dokumentów z elementów niestanowiących ich treść.
- Identyfikacja termów.
- Tworzenie reprezentacji dokumentów w postaci przestrzeni wektorów.

Budowanie modelu odwzorowania dokumentów w klasy

- Budowa modelu przeprowadzana jest zwykle w iteracyjny i interaktywny sposób.
- Kroki konstrukcji modelu:
  - Selekcja cech;
  - Zastosowanie algorytmu klasyfikacji;
  - Ocena jakości klasyfikatora;
  - Weryfikacja parametrów algorytmu.

Testowanie i ocena modelu

- Klasyfikacja dokumentów zbioru testowego.
- Ocena klasyfikacji na bazie rzeczywistych klas dokumentów oraz tych przewidywanych przez klasyfikator.
- Zastosowanie standardowych miar ceny jakości klasyfikacji.

Klasyfikacja nowych dokumentów za pomocą otrzymanego modelu

- Przypisanie nowych dokumentów (tj. dokumentów, których klasy są nieznane) do predefiniowanych klas za pomocą otrzymanego modelu.
- Jakość klasyfikacji może być oceniona na podstawie wiedzy dotyczącej klasyfikowanych dokumentów lub poprzez zastosowanie innego narzędzia, np. dodatkowy klasyfikator.

Gromadzenie danych dotyczących dokumentów – przykład

Gromadzenie danych dotyczących dokumentów – przykład c.d.

Anthropology, "anthropology anthropology anthropology consists of four subfields cultural anthropology physical anthropology archaeology and linguistics beyond these subfields concentrations are offered in biological anthropology and cross cultural comparison the anthropology major provides students with a broad social and behavioral science background and prepares students for a range of careers from public service to marketing and international management through independent study students work closely with faculty doing research students regularly attend professional anthropology meetings and other discipline related events special programs include summer field schools in archaeology internships in applied anthropology and participation in department sponsored diversity training institutes program of study ba department chair michael park location diloreto hall 110 phone 830 2610 department website", **A**

Art, "art art the art department s undergraduate degree program offers a wide range of visual art options including painting printmaking sculpture ceramics and graphic design both concepts and technical excellence are stressed within a curriculum that encourages all forms of creative explorations via a developing professional exchange with faculty all majors must complete a successful portfolio review of at least 10 works to become eligible for upper division courses study plans for students are developed on an individual basis consistent with the goals identified by the student and the advisor the department houses within the samuel chen art center a gallery that offers regular shows for professional exhibits including the works of internationally known artists sol lewitt cleve gray and robert cottingham opportunities abound for internships with community based design firms museums and galleries programs of study ba ms department chair cassandra broadus garcia location maloney hall 151 phone 832 2620 department website", **B**

...

Przetwarzanie danych dotyczących dokumentów – przykład

- Reprezentacja TFIDF:

No.	1: document_name	2: academic	3: accelerator	4: accounting	5: accreditation	...	609: writing	610: year	611: york	612: **document_class**
	Nominal	Numeric	Numeric	Numeric	Numeric		Numeric	Numeric	Numeric	Nominal
1	Anthropology	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
2	Art	0.0	0.0	0.0	0.0		0.0	0.0	0.0	B
3	Biology	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
4	Chemistry	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
5	Communication	1.609438	0.0	0.0	0.0		0.0	0.0	0.0	B
6	Computer	0.0	0.0	0.0	2.995732		0.0	0.0	0.0	A
7	Justice	1.609438	0.0	0.0	0.0		0.0	0.0	0.0	B
8	Economics	1.609438	0.0	2.995732	0.0		0.0	0.0	0.0	A
9	English	0.0	0.0	0.0	0.0		2.995732	0.0	0.0	B
10	Geography	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
11	History	0.0	0.0	0.0	0.0		0.0	0.0	0.0	B
12	Math	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
13	Languages	0.0	0.0	0.0	0.0		0.0	0.0	0.0	B
14	Music	0.0	0.0	0.0	0.0		0.0	0.0	0.0	B
15	Philosophy	1.609438	0.0	0.0	0.0		0.0	0.0	0.0	B
16	Physics	0.0	2.995732	0.0	0.0		0.0	0.0	0.0	A
17	Political	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
18	Psychology	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
19	Sociology	0.0	0.0	0.0	0.0		0.0	0.0	0.0	A
20	Theatre	0.0	0.0	0.0	0.0		0.0	2.995...	2.995...	B

Selekcja cech

- Celem selekcji cech jest znalezienie (jak najmniejszego) podzbioru atrybutów, który najlepiej opisuje zbiór dokumentów ze względu na zadanie klasyfikacji.
- Typowym kryterium przy selekcji cech jest uzyskanie jak największej rozróżnialności klas obiektów.

Selekcja cech na podstawie metody opartej na podobieństwie

- Każdy atrybut ma na starcie przypisaną taką samą wagę.
- Dla każdego dokumentu znajdowana jest z góry ustalona liczba najbliższych sąsiadów z tej samej i innych klas.
- Jeżeli znaleziony sąsiad z tej samej (innej) klasy ma inną wartość danego atrybutu, to waga tego atrybutu jest obniżana (podwyższana).
- Wybierane są atrybuty z najwyższymi wagami.

Przetwarzanie danych dotyczących dokumentów – przykład c.d.

	history	*science*	*research*	*offers*	*students*	*hall*	Class
Anthropology	0	0.537	0.477	0	0.673	0.177	A
Art	0	0	0	0.961	0.195	0.196	B
Biology	0	0.347	0.924	0	0.111	0.112	A
Chemistry	0	0.975	0	0	0.155	0.158	A
Communication	0	0	0	0.780	0.626	0	B
Computer Science	0	0.989	0	0	0.130	0.067	A
Criminal Justice	0	0	0	0	1	0	B
Economics	0	0	1	0	0	0	A
English	0	0	0	0.980	0	0.199	B
Geography	0	0.849	0	0	0.528	0	A
History	0.991	0	0	0.135	0	0	B
Mathematics	0	0.616	0.549	0.490	0.198	0.201	A
Modern Languages	0	0	0	0.928	0	0.373	B
Music	0.970	0	0	0	0.170	0.172	B
Philosophy	0.741	0	0	0.658	0	0.136	B
Physics	0	0	0.894	0	0.315	0.318	A
Political Science	0	0.933	0.348	0	0.062	0.063	A
Psychology	0	0	0.852	0.387	0.313	0.162	A
Sociology	0	0	0.639	0.570	0.459	0.237	A
Theatre	0	0	0	0	0.967	0.254	B

Metoda *k*-najbliższych sąsiadów w klasyfikacji dokumentów

- Dana jest miar odległości dokumentów: podobieństwo cosinusowe.
- Dla każdego dokumentu, który mam być sklasyfikowany liczona jest jego podobieństwo do dokumentów zbioru treningowego.
- Dokumenty zbioru treningowego są sortowane według uzyskanych podobieństw.
- Klasa danego dokumentu jest określa na podstawie wyniku głosowania *k* najbliższych dokumentów.

Podobieństwo dokumentów – przykład

- Klasyfikowany dokumentów: Theatre (B)

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075
Anthropology	A	0.695979
Communication	B	0.605667
Geography	A	0.510589
Sociology	A	0.504672
Physics	A	0.385508
Psychology	A	0.343685
Mathematics	A	0.242155
Art	B	0.238108
Music	B	0.207746
Chemistry	A	0.189681
Computer Science	A	0.142313
Biology	A	0.136097
Modern Languages	B	0.0950206
Political Science	A	0.0762211
English	B	0.0507843
Philosophy	B	0.0345299
History	B	0
Economics	A	0

Dobór parametru *k* – przykład (1)

- Niech k=1.
- Przy zastosowaniu metody 1-NN rozpatrujemy następujące podobieństwo:

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075

- Liczby głosów oddane na poszczególne klasy: n(A)=0, n(B)=1.
- Dokument *Theatre* zostaje przypisany do klasy B, a zatem wynik klasyfikacji jest poprawny.

Dobór parametru *k* – przykład (2)

- Niech k=2.
- Przy zastosowaniu metody 2-NN rozpatrujemy następujące podobieństwa:

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075
Anthropology	A	0.695979

- Liczby głosów oddane na poszczególne klasy: n(A)=1, n(B)=1.
- Dokument *Theatre* nie jest przypisany do żadnej klasy – konflikt klas.

Dobór parametru *k* – przykład (3)

- Niech k=3.
- Przy zastosowaniu metody 3-NN rozpatrujemy następujące podobieństwa:

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075
Anthropology	A	0.695979
Communication	B	0.605667

- Liczby głosów oddane na poszczególne klasy: n(A)=1, n(B)=2.
- Dokument *Theatre* zostaje przypisany do klasy B, a zatem wynik klasyfikacji jest poprawny.

Dobór parametru *k* – przykład (4)

- Niech k=5.
- Przy zastosowaniu metody 5-NN rozpatrujemy następujące podobieństwa:

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075
Anthropology	A	0.695979
Communication	B	0.605667
Geography	A	0.510589
Sociology	A	0.504672

- Liczby głosów oddane na poszczególne klasy: n(A)=3, n(B)=2.
- Dokument *Theatre* zostaje przypisany do klasy A, a zatem wynik klasyfikacji jest niepoprawny.

Metoda k-najbliższych sąsiadów ważonej odległości

- Głosy *k* najbliższych sąsiadów są ważone za pomocą ich odległości (podobieństw) do klasyfikowanych obiektów.
- Wagi w mogą zostać ustalone w jeden z następujących sposób:
  - $w = \mathrm{sim}(d_1, d_2)$ np. $\mathrm{sim}(d_1, d_2) = 3/4$ to $w = 3/4$;
  - $w = 1/[1 - \mathrm{sim}(d_1, d_2)]$ np. $\mathrm{sim}(d_1, d_2) = 3/4$ to $w = 4$;
  - $w = 1/[1 - \mathrm{sim}(d_1, d_2)]^2$ np. $\mathrm{sim}(d_1, d_2) = 3/4$ to $w = 16$.

Metoda k-najbliższych sąsiadów ważonej odległości – przykład

- Niech k = 4 oraz $w = sim(d_1, d_2)$.
- Na podstawie następujących podobieństw

Document	Class	Similarity to Theatre
Criminal Justice	B	0.967075
Anthropology	A	0.695979
Communication	B	0.605667
Geography	A	0.510589

otrzymujemy

$$n(A) = sim(An, Th) + sim(Ge, Th) = 0.695979 + 0.510589 = 1.206568$$

$$n(B) = sim(CJ, Th) + sim(Co, Th) = 0.967075 + 0.605667 = 1.572742$$

zatem dokument *Theatre* zostaje przypisany do klasy B.

Metoda oceny modelu

- Otrzymany model klasyfikacji dokumentów może zostać oceniony przy zastosowaniu jednej ze standardowych metod.
- Wybór metody może być uzależniony od
  - zastosowanego klasyfikatora (np. uczenie z przykładów wyklucza użycie metody ponownego podstawienia).
  - statystycznej charakterystyki dokumentów (np. dla mniej reprezentatywnych/mniejszych danych preferowana jest bardziej metoda krzyżowej walidacji niż metoda trenuj i testuj).

Dokładność klasyfikacji dokumentów – metoda krzyżowej walidacji – przykład

- Liczba testów krzyżowej walidacji: 20.
- Standardowa metoda *k*-nn:

1-NN	3-NN	5-NN	19-NN
1.00	1.00	0.90	0.55

- Metoda *k*-nn ważonej odległości:

1-NN	3-NN	5-NN	19-NN
1.00	1.00	1.00	0.85