

ORACLE®

Oracle Data Mining 10g

Zastosowanie algorytmu
„Support Vector Machines”
do problemów biznesowych

Piotr Hajkowski
Oracle Consulting

Oracle Expert Services

Agenda

- Podstawy teoretyczne algorytmu SVM
 - SVM w bazie danych
 - Klasyfikacja
 - Regresja
 - Wykrywanie anomalii
 - Parametry sterujące budową algorytmu

Agenda

- Praktyczne zastosowania
 - Korzyści biznesowe
 - Zastosowanie SVM w różnych dziedzinach i sektorach gospodarki
 - Przykład: Ocena ryzyka kredytowego w banku
 - Demonstracja narzędzia Oracle Data Miner
 - Stosowanie modelu do nowych danych

Data Mining w bazie danych

- Wzrastające znaczenie technologii analitycznych
 - Duże wolumeny danych muszą być przetworzone / przeanalizowane
 - Współczesne techniki data mining są efektywne i jednocześnie mają wysoką dokładność
- Wyzwania stawiane przez data mining
 - Złożone metodologie
 - Wymagające obliczeniowo

Dlaczego SVM znalazł się w bazie Oracle?

- Bardzo efektywny klasyfikator
- Mocne podstawy teoretyczne
 - Teoria Vapnika-Chervonenkisa (VC)
- Właściwości regularyzacyjne (gładkość)
 - Dobre uogólnianie dla nietypowych danych
- Doskonały algorytm dla „trudnych” danych o wielu zmiennych wejściowych
 - tekst, obrazy, bioinformatyka
- SVM może być zastosowany do tej samej klasy problemów co sieci neuronowe / RBF

Algorytmy w 10gR2

Technika Data Mining

Algorytm

Klasyfikacja

Decision Tree

Naïve Bayes

Support Vector Machine

Adaptive Bayes Network

Regresja

Support Vector Machine

Istotność atrybutów

MDL

(Minimum Description Length)

Reguły współwystępowania

A Priori

Analiza skupień

KMeans

od 10gR2

OCluster

od 10gR1

od 9iR2

Ekstrakcja cech

NMF

(Non-negative Matrix Factorization)

Wykrywanie anomalii

Support Vector Machine

ORACLE

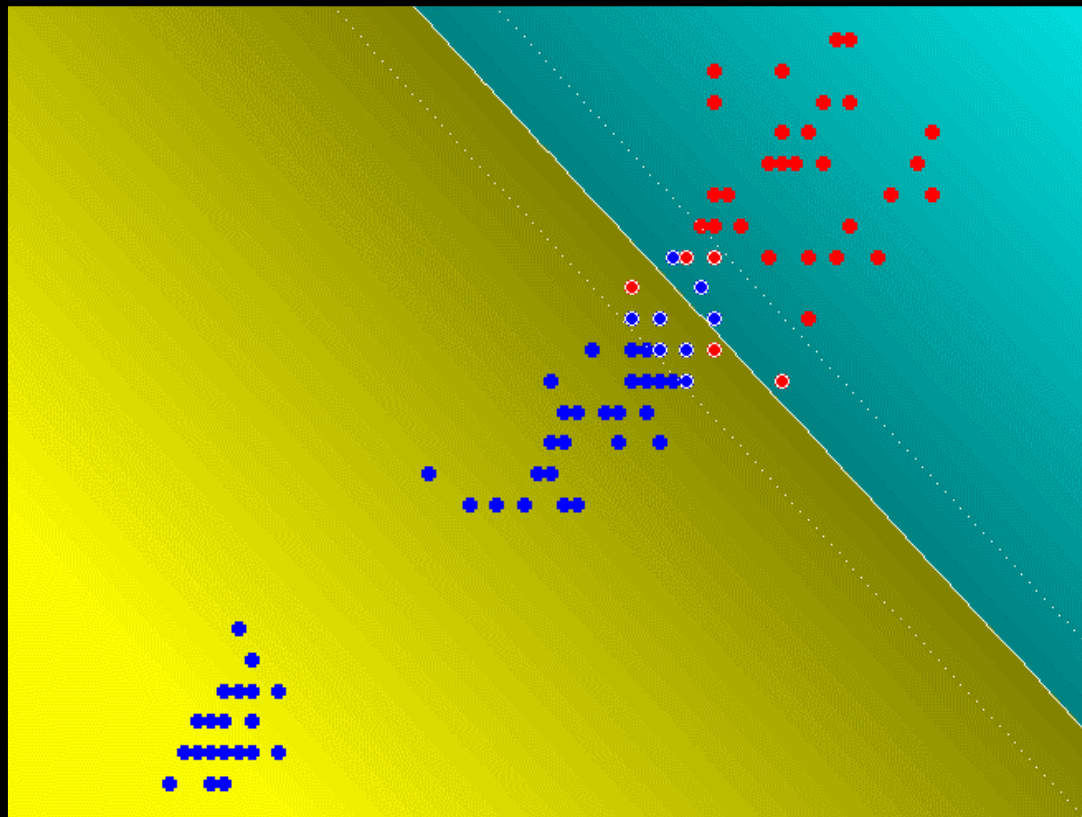
SVM w bazie Oracle

- Oracle Data Mining (ODM)
 - Komercyjna implementacja SVM w bazie wspierająca klasyfikację, regresję i wykrywanie anomalii
 - SVM jest produktem przeznaczonym dla programistów i analityków data mining
 - Sposób implementacji w bazie kładzie nacisk na łatwość użycia i wydajność
- Najważniejsze wyzwania
 - Dobra skalowalność
 - duże wolumeny danych, małe wymagania na pamięć, krótki czas odpowiedzi
 - Wysoka trafność przy domyślnych parametrach

SVM w bazie Oracle

- Typy technik data mining
 - Klasyfikacja: binarna i wieloklasowa
 - Regresja
 - Wykrywanie anomalii (jedna klasa)
- Funkcje jądra (do odwzorowania przestrzeni cech)
 - Funkcje liniowe
 - Funkcje Gaussa

SVM do klasyfikacji: Klasy separowalne



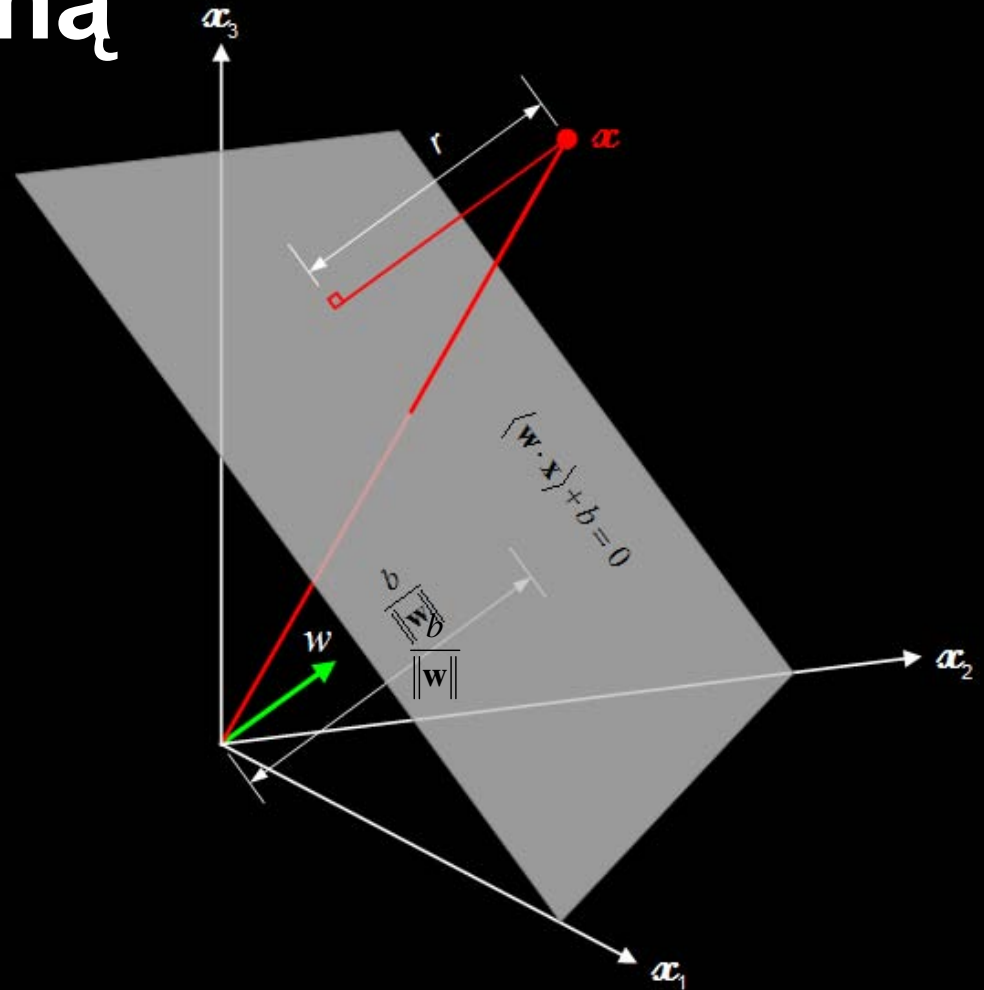
ORACLE

Copyright © 2006 Oracle Corporation

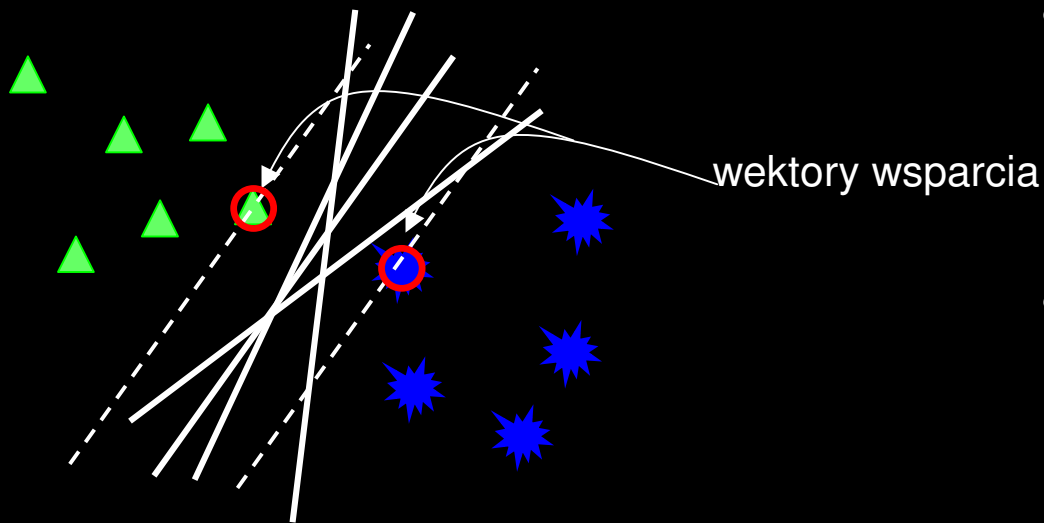
Model SVM jest hiperpłaszczyzną

Model SVM definiuje hiperpłaszczyznę w przestrzeni cech jako wektor (\mathbf{w}) prostopadły do płaszczyzny i przesunięcie (b)

$$f = \text{sign}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b)$$

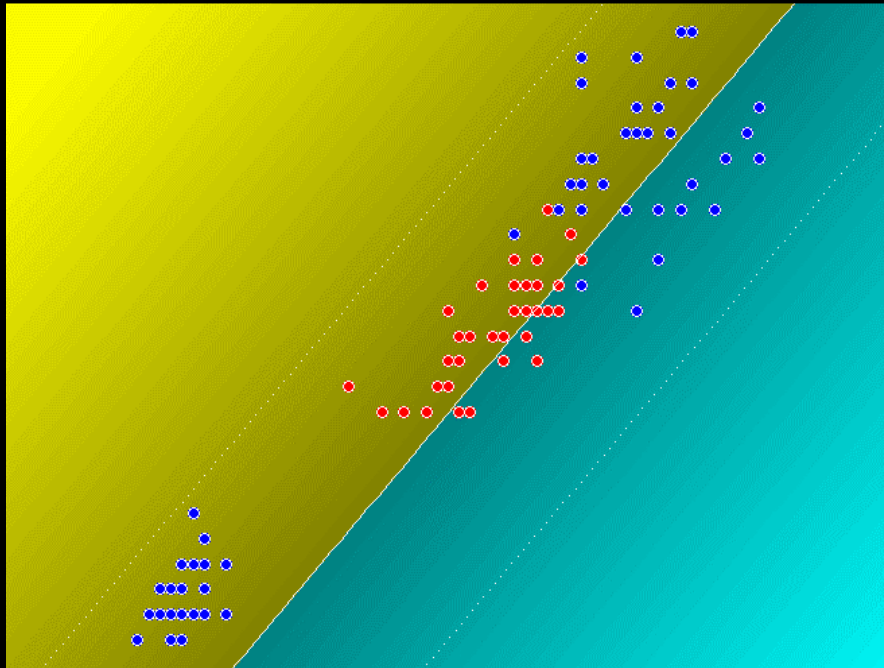


Modele z największym marginesem

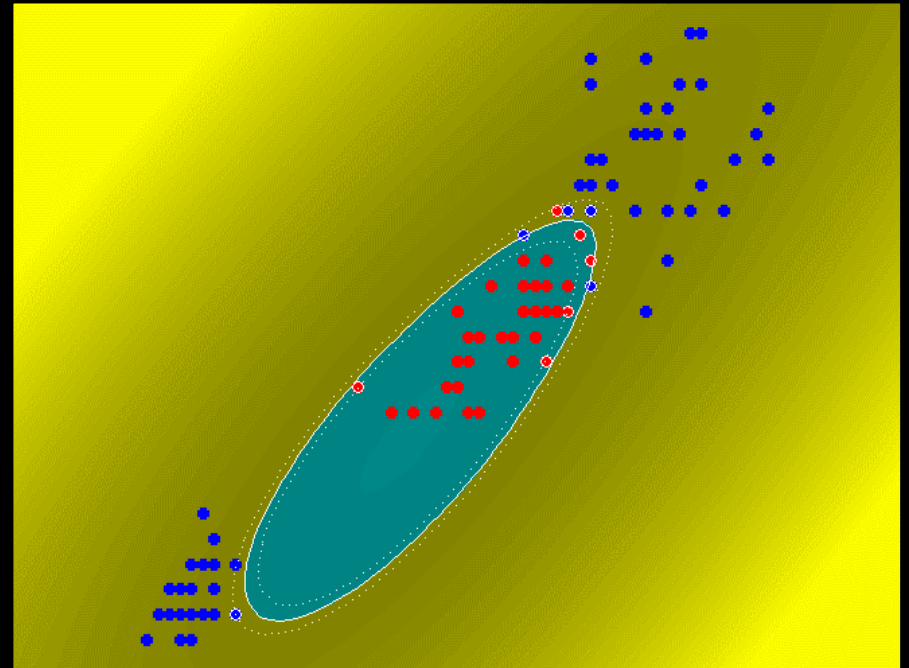


- Wiele różnych modeli separuje klasy
- SVM wybiera płaszczyznę z największą separacją klas

SVM do klasyfikacji: Klasy liniowo nieseparowalne



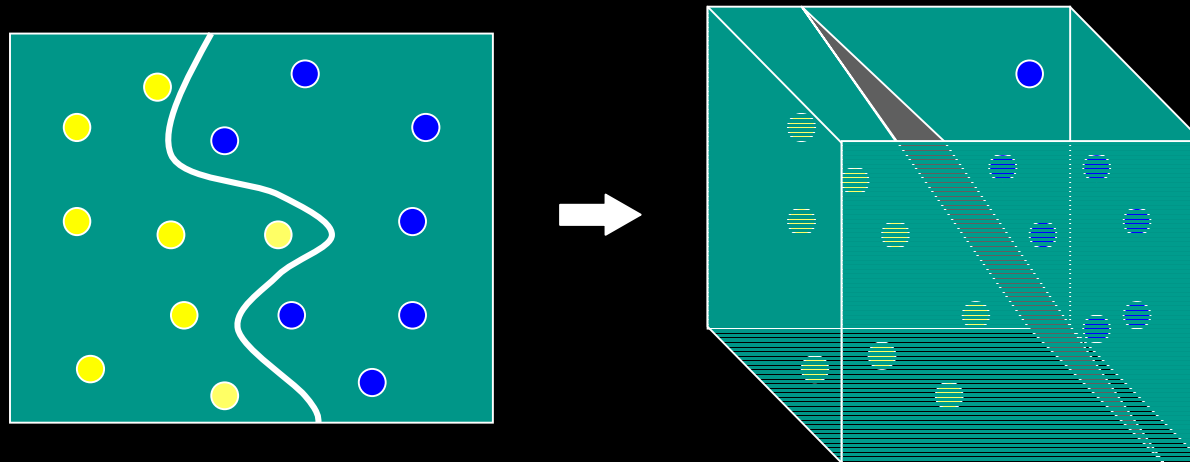
Klasy liniowo
nieseparowalne



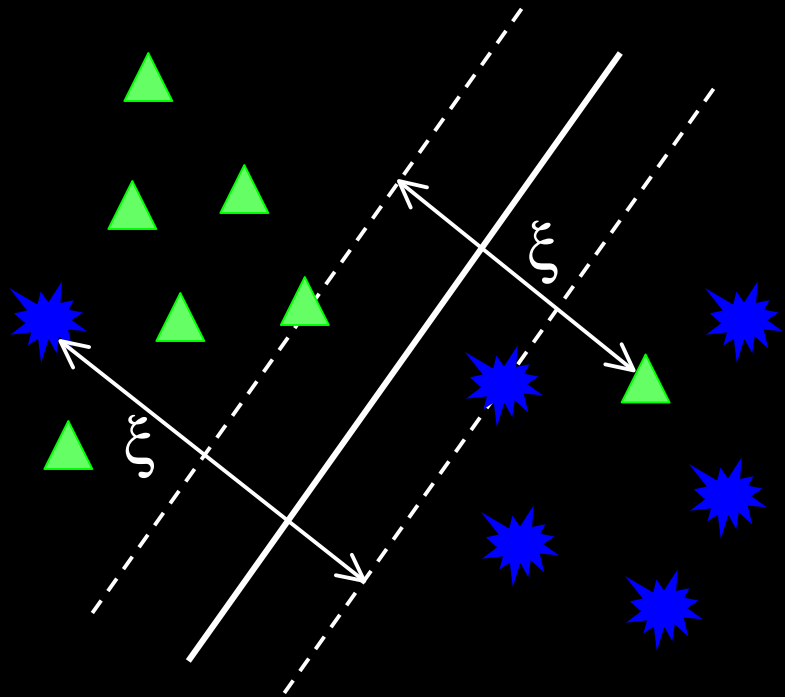
Klasy odseparowane
funkcją nieliniową

Transformacja danych do większej przestrzeni

1. Transformujemy dane przy użyciu nieliniowego odwzorowania (funkcji jądra) do przestrzeni z większą liczbą wymiarów
 - Funkcje jądra: Gaussa i wielomianowe
2. Trenujemy liniowy model w nowej przestrzeni cech

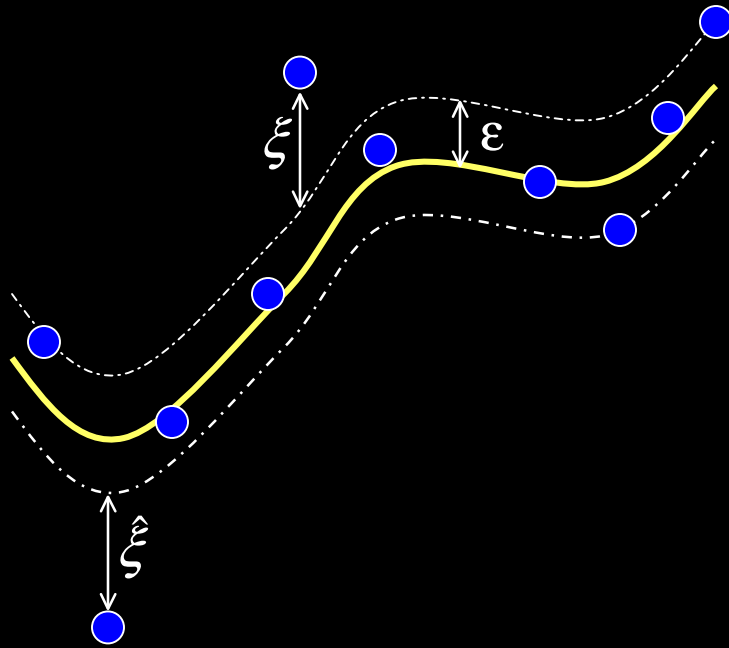


Miękki margines dla danych nieseparowalnych



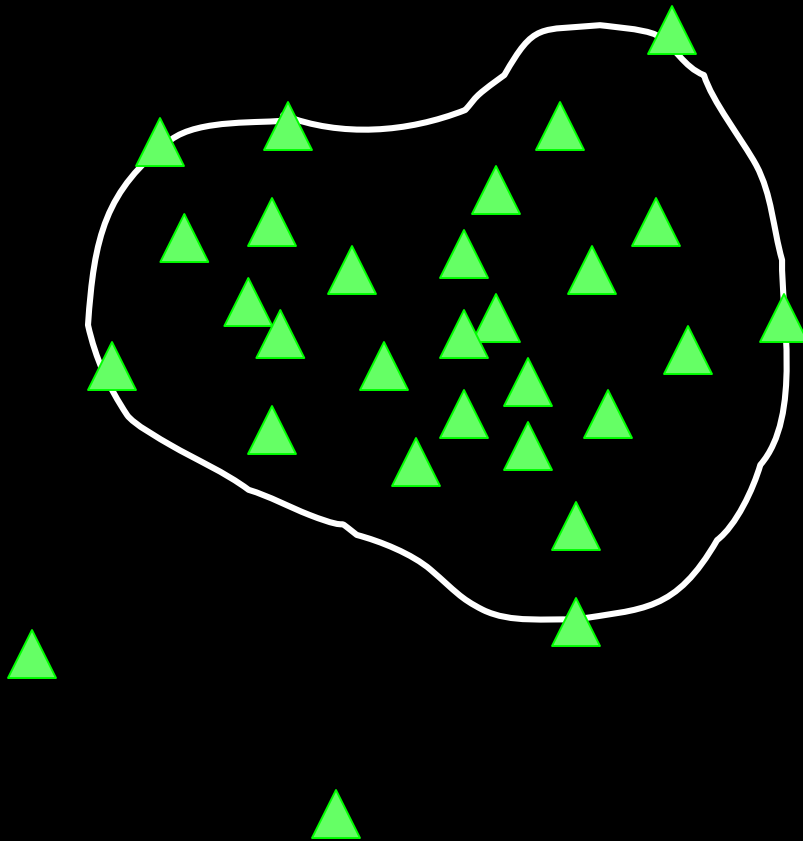
- Współczynnik złożoności pozwala zrównoważyć złożoność modelu i liczby błędów w czasie trenowania modelu
- Wbudowany mechanizm pozwalający uzyskać dobry błąd uogólniania

SVM do regresji



- Epsilon-funkcja niewrażliwej straty
- Parametr epsilon determinuje gładkość dopasowania
- Parametr złożoności C równoważy złożoność modelu i błędy trenowania

SVM do wykrywania anomalii



- Wykrywanie przypadków nietypowych
 - Przypadki typowe vs nietypowe
- Rozróżnienie pomiędzy znaną klasą a światem nieznanymi przypadkami

Dokładność modelu: wpływ użytkownika

Niedoświadczeni użytkownicy mogą uzyskać
wyjątkowo słabe rezultaty

	Dokładność laika	Dokładność eksperta
Fizyka	0.67	0.97
Bioinformatyka	0.57	0.79
Transport	0.02	0.88

Wsparcie Oracle'a w przygotowaniu danych

- Automagiczne przygotowanie danych
 - Usunięcie przypadków nietypowych
 - Normalizacja
 - Przekodowanie atrybutów kategoriycznych do numerycznych
- Wsparcie przez
 - pakiet `dbms_data_mining_transform`
 - Oracle Data Miner

Estymacja „w locie” parametrów dla SVM

- Bazuje na danych (próbki)
- Niski koszt obliczeniowy
- Zapewnia dobre uogólnienie
 - Pozwala uniknąć przetrenowania modelu
 - Model jest zbyt złożony i dane są zapamiętywane
 - Pozwala uniknąć niedotrenowania modelu
 - Model nie jest wystarczająco złożony by odwzorować strukturę danych

SVM: ustawienia dla klasyfikacji

Parametr	Stosuje się do:		Domyśl.	Zakres
Kernel type	Linear	Gauss	Linear	Linear, Gauss
Error tolerance	✓	✓	0.001	(0, 0.1)
Complexity	✓	✓	Est	> 0
Standard dev	—	✓	Est	> 0
Cache size	—	✓	50M	> 0

SVM: ustawienia dla regresji

Parametr	Stosuje się do:		Domyśl.	Zakres
	Linear	Gauss		
Kernel function	Linear	Gauss	Linear	Linear, Gauss
Error tolerance	✓	✓	0.001	(0, 0.1)
Complexity	✓	✓	Est	> 0
Standard dev	—	✓	Est	> 0
Cache size	—	✓	50M	> 0
Epsilon	✓	✓	Est	> 0

Typy funkcji jądra

- Linear (domyślnie)
 - Problemy z dużą liczbą zmiennych wejściowych
 - Problemy liniowo separowalne
 - Jest w stanie otrzymać współczynniki przez `get_model_details_svm` (PL/SQL API)
 - małe, szybkie modele
- Gaussa
 - Problemy z małą liczbą zmiennych wejściowych
 - Problemy liniowo nie separowalne
 - Modele mogą być duże i wolne

Tolerancja błędów i pamięć

- Tolerancja błędów (Error tolerance)
 - Precyzja dopasowania modelu do danych
 - 0.001 (domyślnie)
 - Większa tolerancja przyspieszy trenowanie, jednakże może dawać słabą precyzję
- Wielkość pamięci podręcznej (Cache size)
 - Miejsce na funkcje jądra dla funkcji Gaussa
 - 50MB (domyślnie)
 - Zwiększenie pamięci przyspieszy trenowanie, jednakże może zużyć jej zbyt dużo – rozważyć zwiększenie PGA

Współczynnik złożoności

- Współczynnik złożoności
 - Reguluje równowagę między złożonością modelu a liczbą błędów trenowania modelu
 - Zwiększenie złożoności skutkuje mniejszą liczbą błędów trenowania, jednakże może zaburzyć uogólnienie modelu → przetrenowanie vs. niedopasowanie
 - Estymowany automatycznie

Odchylenie standardowe

- Odchylenie standardowe dla funkcji Gaussa
 - Zwiększenie standardowego odchylenia tworzy gładsze granice decyzyjne, jednakże może skutkować niedopasowaniem
 - Estymowany automatycznie

SVM i wagi

- Wagi (priors) pozwalają ustawić różne koszty kar w modelu SVM dla każdej z klas

$$\frac{C_1}{C_2} \sim \frac{p_1}{p_2}$$

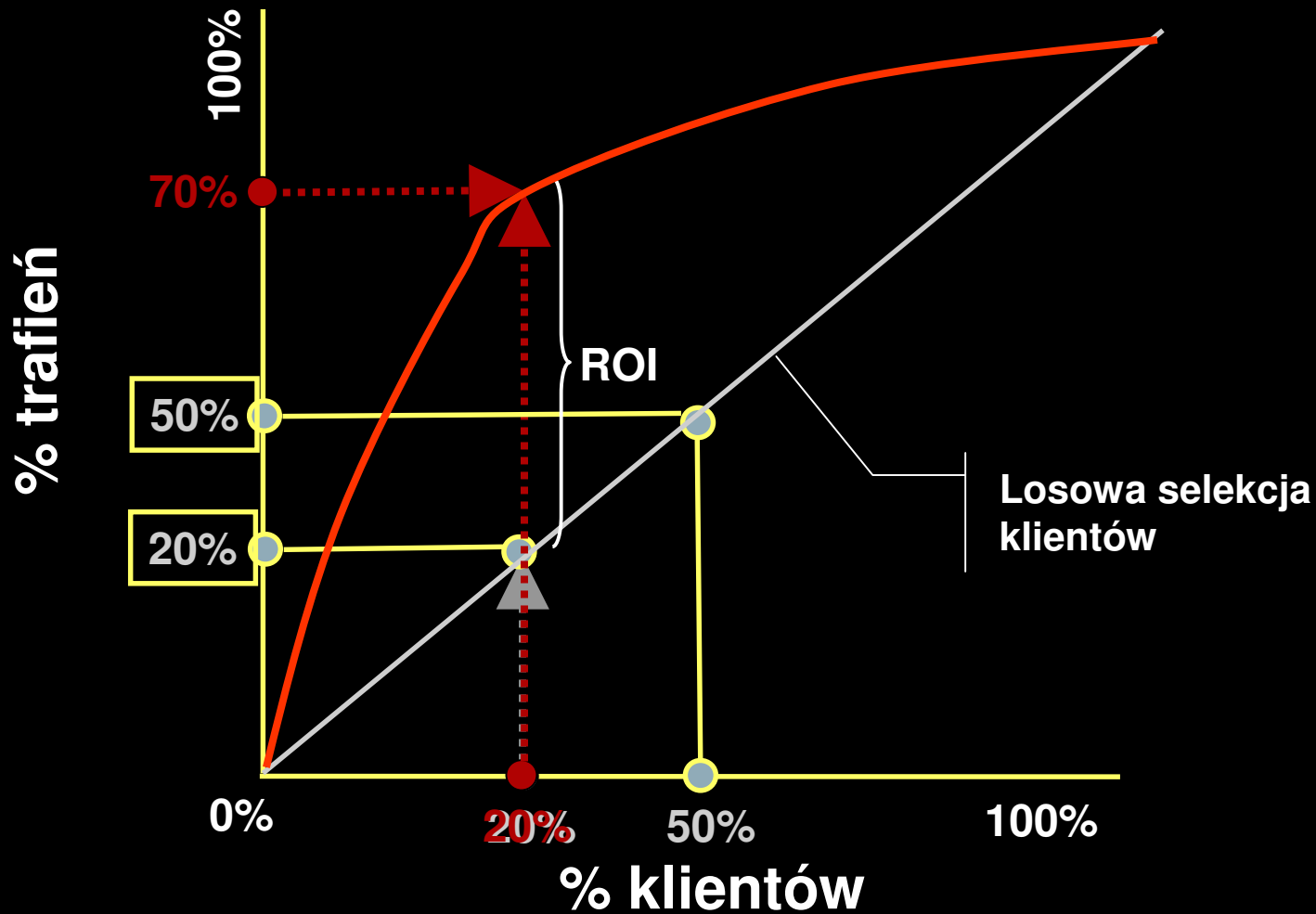
Epsilon

- Epsilon
 - Określa szerokość tunelu wolnego od błędów
 - Kontroluje gładkość dopasowania – większy epsilon daje gładszą funkcję
 - Wpływa na liczbę wektorów wsparcia – większy epsilon daje mniejsze modele (z mniejszą liczbą wektorów wsparcia) ale prawdopodobnie mniej dokładne
 - Estymowany automatycznie

Agenda

- Praktyczne zastosowania
 - Korzyści biznesowe
 - Zastosowanie SVM w różnych dziedzinach i sektorach gospodarki
 - Przykład: Ocena ryzyka kredytowego w banku
 - Demonstracja narzędzia Oracle Data Miner
 - Stosowanie modelu do nowych danych

Wartość Data miningu dla biznesu



ORACLE

Copyright © 2006 Oracle Corporation

Typowe biznesowe zastosowania data miningu i rezultaty

- Mailing bezpośredni ➤ Docelowe listy mailingowe (KB)
 - Analiza przejść do konkurencji ➤ Lista klientów o wysokim prawdopodobieństwie odejścia (KB)
 - Analiza ryzyka kredytowego ➤ Indywidualny rating kredytowy (KW)
 - Wykrywanie nadużyć ➤ Oflagowanie transakcji do kontroli (WA)
 - Systemy bezpieczeństwa ➤ Oflagowanie podejrzanej operacje/zdarzenia (WA)
-
- KB – klasyfikacja binarna: tak, nie
 - KW – Klasyfikacja wieloklasowa: A-, A, A+, B-, B, B+
 - WA – Wykrywanie anomalii: inny niż 95% przypadków

Cykl życia projektu data mining'owego

- 6 faz projektu data mining'owego
 - Zrozumienie biznesu
 - Identyfikacja danych opisujących biznes
 - Przygotowanie danych
 - Budowa modelu
 - Ocena jakości modelu
 - Zastosowanie modelu w praktyce

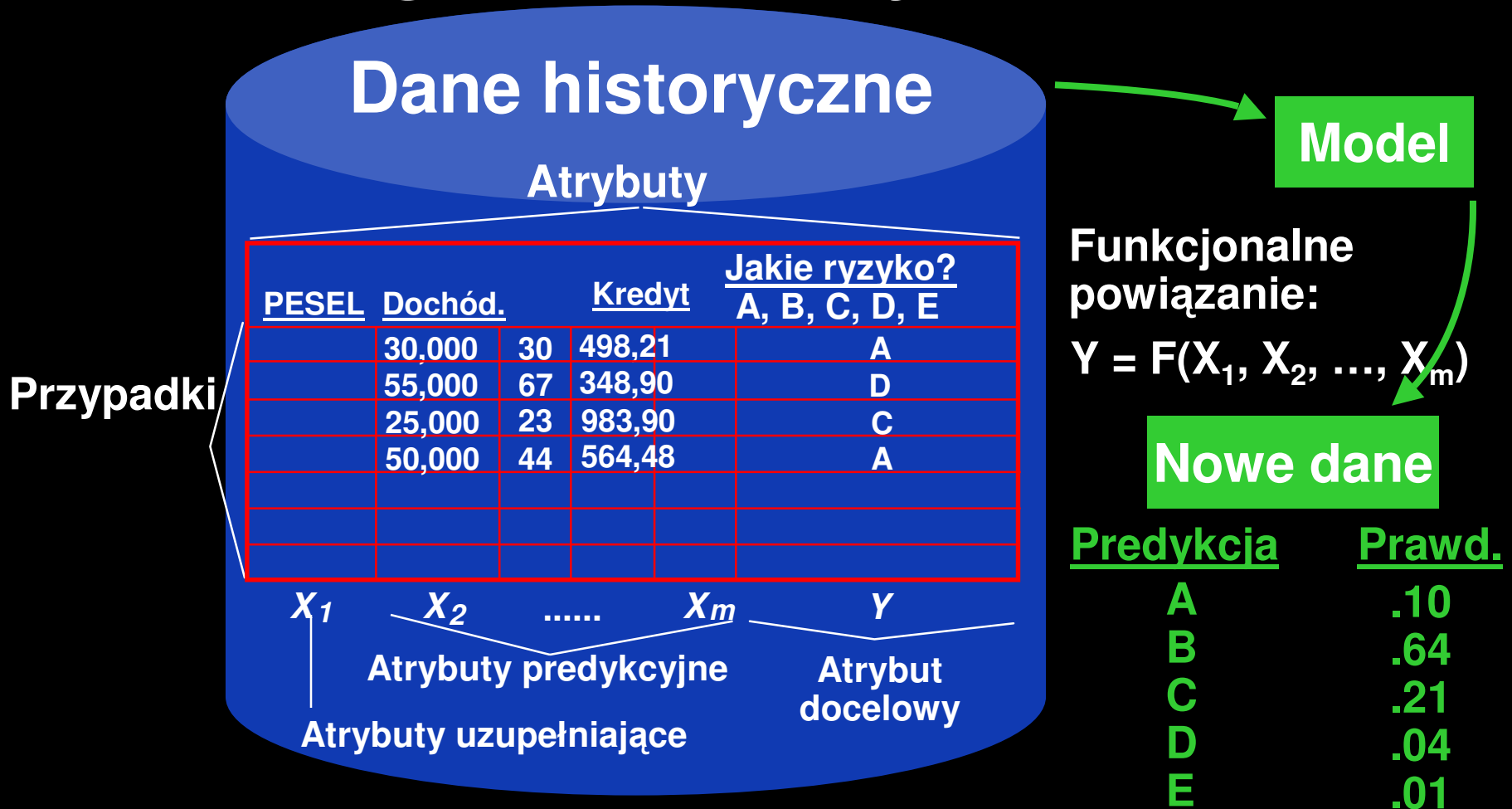
Przykład: Ocena ryzyka kredytowego w banku

- Ryzyko kredytowe określa prawdopodobieństwo, że klient nie spłaci zaciągniętego kredytu
- Znając ryzyko bank może podjąć działania zabezpieczające
- Jeśli założymy, że klienci nie spłacający kredytów w przeszłości są podobni to analogicznych klientów w przyszłości, to możemy poszukać podobieństw lub ukrytych wzorów zachowań

Przykład: Postawienie problemu w kategoriach danych

- Oznaczmy ryzyko na 5 stopniowej skali od A do E:
 - A – najmniejsze ryzyko
 - E – największe ryzyko
- Cechy kredytobiorcy
 - wysokość kredytu, dochód,
 - stan cywilny, płeć, wiek,
 - ...
- Dane historyczne
 - Kilka poprzednich lat

Przykład: Postawienie problemu w kategoriach danych



Oracle Data Miner 10g

- Przygotowanie danych
- Budowa modelu
- Ocena jakości modelu

Scoring: Predykcja w SQL

- Scoring jako zintegrowana część aplikacji bazodanowej
- Możliwość użycia scoringu w SQL jak każdej innej funkcji
- Przetwarzanie strumieniowe
 - Scoring w czasie wykonania zapytania
 - Możliwych jest wiele modeli w jednym zapytaniu
 - Pierwsze rezultaty są szybko zwracane

Scoring: przykład z naszym modelem ryzyka kredytowego

Wybierz klientów dla których prawdopodobieństwo ryzyka kredytowego 'A' jest większe 20%

```
select *  
from credit_risk_customers_t t  
where predition_probability(  
    credit_risk_c89015_sv, 'A' using *) > 0.2
```

Scoring: Predykcja w SQL

Ile otrzymaliśmy przychodu w ostatnim roku od klientów, którzy mieszkają w Warszawie i prawdopodobnie uciekną do konkurencji?

```
Select sum(s.amount_sold)
  from sales s, customers c
 where s.cust_id = c.cust_id
 and c.city = 'Warszawa'
 and prediction (ModelUcieczek using c.*)
 = 'Ucieknie';
```

ORACLE®