UAI Universitas Al Azhar Indonesia

UJIAN AKHIR SEMESTER GENAP
TAHUN AKADEMIK 2016/2017
FAKULTAS SAINS DAN TEKNOLOGI

YAYASAN PESANTREN ISLAM AL-AZHAR

Jurusan	:	Teknik Informatika	**Hari/Tanggal**	:	Selasa, 18 Juli 2017
Mata Kuliah	:	Data Mining	**Sifat**	:	Open book, kalkulator
Nama Dosen	:	Ir. Endang Ripmiatin, M.T. Ali Akbar S., S.T., M.Sc.	**Waktu**	:	150 menit

Peraturan

- Jawab semua soal berikut
- Notasi pemisah ribuan adalah koma (,), sedangkan desimal ditulis dengan titik (.)

1 Supervised Learning

1.1 Model Linear dan Optimasi

(a) Gambarkan metode optimasi numerik dengan *gradient descent* untuk fungsi error $E(w) = 2w^2$. Deskripsikan cara kerjanya, berikan contoh dalam dua *epoch*, dan tunjukkan dalam gambar tersebut efek besarnya laju pembelajaran (*learning rate*; η). [4 poin]

(b) Jelaskan konsep *underfitting* dan *overfitting* pada kasus regresi! Anda dapat menggunakan gambar untuk membantu penjelasan Anda. [4 poin]

(c) Mengapa kita melakukan regularisasi pada regresi linear? Berikan salah satu contoh metode regularisasi dan dampaknya pada model yang dihasilkan! [4 poin]

(d) Diberikan data latih dengan dua atribut x_1 dan x_2 seperti pada Gambar 1. Apakah ada vektor bobot (*weight vector*) yang dapat menghasilkan *linear classifier* yang secara sempurna dapat membagi kedua kelas? Jika ya, gambarkan batas keputusan dan vektor bobotnya. Jika tidak, jelaskan mengapa tidak ada. [3 poin]

1.2 k-Nearest Neighbours

Diberikan data dengan tiga atribut numerik (x_1, x_2, x_3) sebagai berikut:

```
A:(2,-1,-2) B:(1,3,-3) C:(4,2,0)
D:(3,-1,-1) E:(3,1,-1) F:(5,1,1)
```

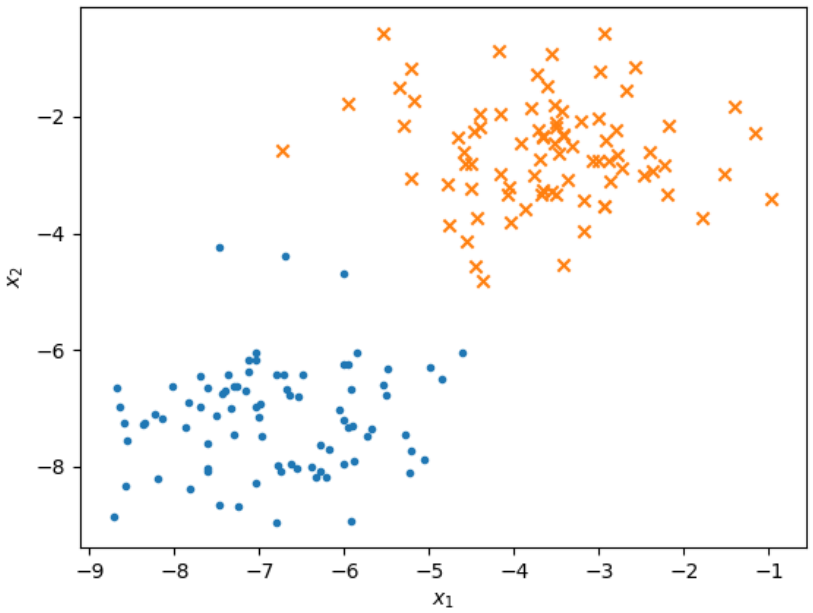
(a) Untuk kasus regresi dengan k-NN, deskripsikan cara memilih nilai k dengan benar. *Metrics* apa yang akan dioptimasi? [4 poin]

(b) Lakukan regresi dengan k-NN untuk memprediksi nilai x_3 jika diberikan $x_1 = 4$ dan $x_2 = 0$. Gunakan $k = 3$ dan Euclidean distance. [4 poin]

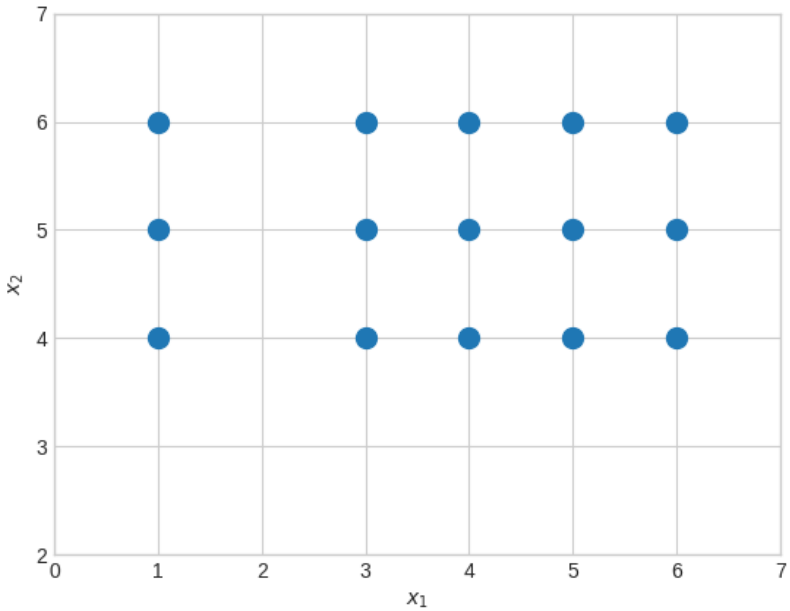
(c) Dalam regresi dengan k-NN, apakah kita akan menemukan kasus seri sehingga memerlukan *tie breaking* seperti pada kasus klasifikasi? Jika ya, apa yang harus dilakukan? Jika tidak, mengapa tidak diperlukan *tie breaking*? [2 poin]

Euclidean distance didefinisikan sebagai

$$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



Gambar 1: Data latih *supervised learning*



Gambar 2: Dataset untuk *unsupervised learning*

2 Unsupervised Learning

Diberikan dataset dengan dua atribut x_1 dan x_2 seperti pada Gambar 2. Dataset tersebut terdiri dari 15 objek yang didefinisikan sebagai berikut:

```
A:(1,4) B:(3,4) C:(4,4) D:(5,4) E:(6,4)
F:(1,5) G:(3,5) H:(4,5) I:(5,5) J:(6,5)
K:(1,6) L:(3,6) M:(4,6) N:(5,6) O:(6,6)
```

2.1 Clustering

(a) Jalankan algoritma k-Means untuk dataset pada Gambar 2. Gunakan $k = 2$ dan Euclidean distance. Inisialisasi nilai $\mu_1 = (1, 3)$ dan $\mu_2 = (6, 3)$. Tunjukkan proses *clustering* yang Anda lakukan. Laporkan nilai akhir μ_1 dan μ_2 dan tulis daftar objek yang ada dalam masing-masing klaster setelah algoritmanya berhenti. Catatan: Anda tidak perlu menghitung nilai pasti dari jarak Euclidean distance, gunakan pemahaman visual Anda. [5 poin]

(b) Lakukan *agglomerative clustering* untuk data di atas. Gunakan Euclidean distance dan *complete link* untuk penggabungan klaster, dan urutan abjad untuk *tie breaking*. Gambarkan dendogram yang dihasilkan. Catatan: Anda tidak perlu menghitung nilai pasti dari jarak Euclidean distance, gunakan pemahaman visual Anda. [5 poin]

2.2 PCA

(a) Dengan menggunakan ide bahwa vektor eigen akan selalu mengarah ke variansi terbesar, tentukan kemungkinan arah dari komponen prinsipil pertama dari data pada Gambar 2. [2 poin]

(b) Jika diketahui bahwa nilai $\lambda_1 = 2.96$ dan $\lambda_2 = 0.67$, berapa persen variansi yang dijelaskan oleh komponen prinsipil pertama? [3 poin]

3 Sistem Rekomendasi

Tiga orang pengguna sebuah situs basis data film memberikan nilai untuk empat film dengan skala 1-5 bintang sesuai Tabel 1.

Tabel 1: Tabel nilai

Pengguna	Film	Nilai
A	w	4
A	x	5
A	z	5
B	x	3
B	y	4
B	z	3
C	w	2
C	y	1
C	z	3

(a) Buatlah *utility matrix* dari tabel tersebut. [2 poin]

(b) Ubahlah nilai dalam *utility matrix* tersebut sebagai *boolean*, i.e. 1 untuk nilai ≥ 3, 0 untuk < 3. Lalu, hitung Jaccard *similarity* antarpengguna. [3 poin]

(c) Menggunakan nilai asli, normalisasi nilainya berdasarkan rata-rata untuk tiap pengguna, lalu hitung *cosine similarity* antarpengguna. [3 poin]

(d) Berdasarkan (b) dan (c), serta menggunakan satu tetangga terdekat, berapa nilai yang akan diberikan B untuk film w? [2 poin]