UJIAN AKHIR SEMESTER GANJIL
TAHUN AKADEMIK 2017/2018
FAKULTAS SAINS DAN TEKNOLOGI

Jurusan	:	Teknik Informatika	**Hari/Tanggal**	:	Rabu, 24 Januari 2018
Mata Kuliah	:	Data Mining	**Sifat**	:	Buka buku, kalkulator
Nama Dosen	:	Ir. Endang Ripmiatin, M.T. Ali Akbar S., S.T., M.Sc.	**Waktu**	:	120 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

Diberikan data latih seperti pada Tabel 1. Dalam kasus ini, x_1 dan x_2 adalah fitur dengan nilai riil, dan label $y \in \{0, 1\}$.

Tabel 1: Data latih

x_1	x_2	y
0.0	0.0	1
2.0	2.0	1
0.0	2.0	0
2.0	0.0	0

(a) Gambarkan titik-titik data latih Anda dalam grafik dengan sumbu x_1 dan x_2. Berikan tanda titik mana yang merupakan kelas 1 dan kelas 0. [2 poin]

(b) Ingat kembali bahwa model regresi logistik didefinisikan sebagai

$$p(y = 1|\mathbf{x}) = \sigma(w_0 + w_1x_1 + w_2x_2),$$

dengan σ adalah fungsi sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$ dan $\mathbf{w} = (w_0, w_1, w_2)$ adalah parameter dari model. Jika diberikan vektor bobot $\mathbf{w} = (1.0, -2.0, 2.0)$, hitung probabilitas tiap contoh dalam data latih di atas mempunyai label 1 berdasarkan model regresi logistik. [4 poin]

(c) Berapa akurasi dari model yang dihasilkan pada data latih? Akurasi didefinisikan sebagai jumlah data yang diprediksi kelas atau labelnya dengan benar dibagi dengan jumlah data. Jika menggunakan intuisi Anda, apakah ada vektor bobot yang bisa menghasilkan klasifikasi dengan akurasi yang lebih baik? Jika ada, berikan nilai vektor bobotnya. Jika tidak, jelaskan mengapa tidak ada. [4 poin]

(d) Dengan metode *batch gradient descent*, coba perbaiki nilai w_0, w_1, dan w_2 dalam satu *epoch*. Perhatikan bahwa nilai baru dari w_j didefinisikan sebagai:

$$w_j \leftarrow w_j - \eta \sum_{i=1}^{N} (\hat{y}_i - y_i)x_j$$

untuk $j > 0$ dan

$$w_j \leftarrow w_j - \eta \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

untuk $j = 0$. Tunjukkan efek dari perbedaan nilai η yang Anda gunakan. [5 poin]

1.2 k-Nearest Neighbours

Anda akan mengaplikasikan algoritma k-Nearest Neighbours (k-NN) untuk kasus deteksi spam. Data latih Anda adalah e-mail (teks berupa string) yang dilabeli "spam" atau "ham".

(a) Deskripsikan secara detail bagaimana cara kerja algoritma k-NN. [3 poin]

(b) Bagaimana Anda akan merepresentasikan e-mailnya? Tuliskan secara spesifik atribut yang akan digunakan dan apa saja nilai yang mungkin. [3 poin]

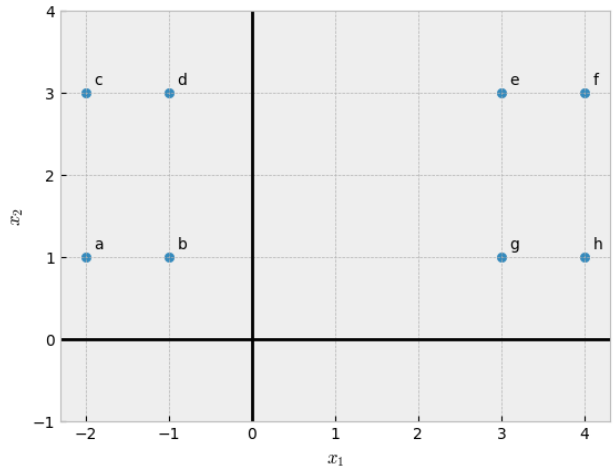
(c) Fungsi jarak apa yang akan Anda gunakan? Tuliskan rumusnya. [2 poin]

(e) K-D trees terkadang digunakan untuk mempercepat proses k-NN. Apakah K-D trees akan cocok pada kasus ini? Jelaskan jawaban Anda. [2 poin]

2 Unsupervised Learning

2.1 Clustering

Misalkan Anda diberikan dataset dengan tiap objek direpresentasikan dengan dua nilai atribut seperti pada Gambar 1.



Gambar 1: *Data points* untuk *clustering*

(a) Jalankan algoritma k-Means untuk dataset pada Gambar 1. Gunakan $k = 2$ dan urutan abjad untuk *tie breaking*. Inisialisasi nilai $\mu_1 = (0, 7)$ dan $\mu_2 = (2, -2)$. 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. [5 poin]

(b) Bagaimana cara menentukan nilai k dalam algoritma k-Means? [2 poin]

(c) Berikan contoh inisialisasi nilai μ_1 dan μ_2 yang dapat menyebabkan konvergensi ke minimum lokal. Tunjukkan pula nilai μ_1 dan μ_2 saat konvergensi tersebut tercapai. [3 poin]

(d) Lakukan *agglomerative clustering* untuk data di atas. Gunakan *complete link* untuk penggabungan klaster hingga hanya tersisa dua klaster. Jika terdapat seri, selesaikan dengan urutan abjad. Gambarkan dendogramnya. Tulis daftar objek yang ada pada masing-masing klaster saat Anda menggunakan dua klaster dari dendogram yang dihasilkan. [5 poin]

2.2 PCA

(a) Dengan menggunakan ide bahwa vektor eigen akan selalu mengarah ke variansi terbesar, tentukan kemungkinan arah dari komponen prinsipil yang dihasilkan dari data pada Gambar 1. [2 poin]

(b) Jika diketahui bahwa nilai $\lambda_1 = 7.43$ dan $\lambda_2 = 1.14$, berapa persen variansi yang dijelaskan oleh komponen prinsipil pertama? [3 poin]