UAI Universitas Al Azhar Indonesia

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

YAYASAN PESANTREN ISLAM AL-AZHAR

Jurusan	:	Teknik Informatika	**Hari/Tanggal**	:	Selasa, 18 April 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 Analisis dan Representasi Data

Sebuah LSM yang bergerak di bidang kesehatan masyarakat mempekerjakan beberapa orang petugas lapangan untuk melakukan survey ke masyarakat untuk mengetahui persebaran kasus tuberkulosis (TB) di masyarakat. Dari survey tersebut, didapatkan data dari 8,732 orang.

(a) Gambar 1 menunjukkan distribusi dari hasil survey tinggi dan berat badan yang dilakukan oleh petugas lapangan. Garis putus-putus pada grafik tersebut merupakan distribusi normal/Gaussian yang dicocokkan dengan data tersebut. Apa yang dapat Anda analisis dari kedua data tersebut? Apakah ada keanehan? Apa yang mungkin menyebabkan hal tersebut? [4 poin]

(b) Batuk berdarah adalah salah satu gejala dari TB. Anda menemukan bahwa 7,758 orang tidak menderita batuk berdarah, 646 menderita batuk berdarah. Artinya, masih ada 328 orang yang tidak teridentifikasi oleh petugas lapangan apakah orang tersebut batuk berdarah atau tidak. Apa yang sebaiknya Anda lakukan kepada 328 data ini? [3 poin]

(c) Masalah apa saja yang dalam ekspektasi Anda akan ditemukan dalam dataset yang diberikan selain yang telah disebutkan di atas? Berikan contohnya! [3 poin]

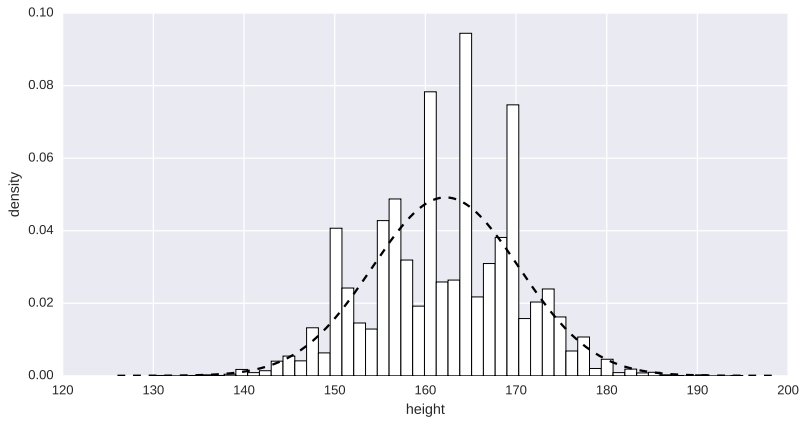
(d) Dari data yang Anda punya, Anda ingin membandingkannya dengan data dari Dinas Kesehatan untuk mengetahui apakah ada orang-orang yang selama ini belum pernah masuk ke data pasien TB milik dinas. Dengan melakukan hal tersebut, harapannya LSM tersebut punya nilai tambah karena telah menemukan kasus-kasus baru di masyarakat. Apa yang dapat Anda lakukan untuk menemukan kandidat kasus-kasus baru tersebut? [4 poin]

LSM tersebut juga mengumpulkan data lain berupa foto hasil pemeriksaan sampel sputum (dahak) di bawah mikroskop. Sampel sputum tersebut telah melalui proses pewarnaan sehingga bakteri tuberkulosis dapat terlihat berupa basil (batang) berwarna merah dengan latar belakangnya akan berwarna kuning muda dengan gradasi hingga warna biru. Beberapa gambar juga mengandung warna hitam jika terdapat benda-benda asing yang terdapat dalam sampel. Anda diminta untuk melakukan klasifikasi untuk menentukan apakah sampel tersebut bernilai TB+ atau TB-. Foto tersebut dibagi dalam 100 lapangan pandang berbentuk *grid* berukuran 10×10. Untuk setiap lapangan pandang, kita dapat menghitung rata-rata nilai tiap pixel-nya. Nilai rata-rata tersebut kemudian didiskritkan menjadi: “merah”, “kuning”, “biru”, “ungu”, dan “hitam”.

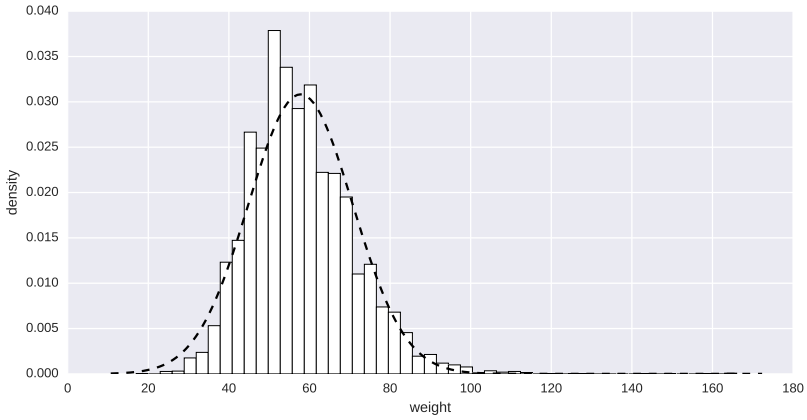
(e) Bagaimana Anda akan merepresentasikan data ini dalam pasangan atribut-nilai? [2 poin]

(f) Ada berapa atribut yang akan kita gunakan? Apakah nilainya nominal, ordinal, atau numerik? [2 poin]

(g) Berapa nilai yang mungkin dari atribut-atribut tersebut? [2 poin]



(a) Distribusi tinggi badan



(b) Distribusi berat badan

Gambar 1: Tinggi dan berat badan dari hasil survey

2 Naïve Bayes

Anda sedang membuat sebuah *classifier* untuk mendeteksi komentar spam di Instagram dengan menggunakan metode Naïve Bayes. Diberikan data seperti pada Tabel 1 dengan `log_char` adalah logaritma dari jumlah karakter komentar dan `has_pattern_yu+k` merupakan kecocokan *regular expression* (regex) "yu+k" dalam teks komentar tersebut. Asumsikan bahwa `log_char` terdistribusi secara Gaussian dan `has_pattern_yu+k` mengikuti distribusi Bernoulli. *Probability mass function* (PMF) dari distribusi Bernoulli adalah:

$$f(x;\theta) = \begin{cases} \theta & ,x = 1 \\ 1-\theta & ,x = 0 \end{cases}$$

sedangkan *probability density function* (PDF) dari distribusi Gaussian adalah:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Tabel 1: Data latih filter komentar spam

log_char	has_pattern_yu+k	label
5.39	True	spam
5.85	False	spam
7.05	False	spam
4.02	True	spam
5.19	False	spam
4.65	False	ham
3.43	False	ham
5.26	True	ham
3.40	False	ham
3.47	False	ham

(a) Hitunglah semua parameter distribusi Gaussian dan Bernoulli yang diperlukan! Tunjukkan perhitungan Anda! [5 poin]

(b) Menggunakan model yang telah terbentuk di bagian (a), klasifikasikan data dengan atribut `log_char` = 3.43 dan `has_pattern_yu+k` = False! Tunjukkan perhitungan Anda! [3 poin]

(c) Anda mendapatkan sebuah data komentar baru yang *mengandung* regex "yu+k". Namun, sayangnya data yang Anda peroleh ini rusak sehingga Anda tidak tahu pasti berapa panjang dari komentar tersebut (tidak ada nilai `log_char` yang dapat dipakai). Apakah Anda dapat mengklasifikasikan data baru ini dengan model yang dibentuk di bagian (a)? Jika ya, tunjukkan proses klasifikasinya. Jika tidak, jelaskan mengapa Anda tidak bisa melakukan klasifikasi tersebut. [2 poin]

(d) Apa yang dimaksud sebagai asumsi "naif" dari Naïve Bayes? Hubungkan penjelasan Anda dengan kasus yang diberikan. [2 poin]

(e) Seandainya semua atribut yang kita pakai terdistribusi Gaussian, apa yang dapat kita lakukan untuk melunakkan asumsi "naif" dari Naïve Bayes? [1 poin]

(f) Mengapa metode Naïve Bayes disebut menghasilkan model generatif? Apa bedanya dengan model yang diskriminatif? [2 poin]

3 Decision Trees dan ROC Curves

3.1 Decision Trees

Terdapat dua variabel, x_1 dan x_2, bernilai 0 atau 1 yang akan Anda gunakan dalam proses klasifikasi Anda. Tujuan dari klasifikasi tersebut adalah menghasilkan model yang dapat merepresentasikan aturan berikut:

$$(x_1 = 1\ AND\ x_2 = 0)\ OR\ (x_1 = 0\ AND\ x_2 = 1)$$

(a) Salah satu algoritma *decision trees* yang dapat digunakan adalah dengan ID3, yaitu dengan melakukan pembagian dengan menghitung nilai *information gain*. Tuliskan *pseudocode* dari algoritma ID3! [3 poin]

(b) Gambarkan pohon klasifikasi yang memenuhi aturan di atas. Catatan: Anda tidak perlu menghitung nilai *information gain*, gunakan intuisi saja. [4 poin]

(c) Deskripsikan mekanisme yang dapat digunakan untuk menghindari *overfitting* dalam *decision trees*! [3 poin]

3.2 ROC Curves

Asumsikan Anda mempunyai 4 contoh dengan kelas positif dan 8 contoh dengan kelas negatif. Anggaplah bahwa Anda menggunakan klasifikasi yang menghasilkan nilai probabilistik $p(+|\mathbf{x})$. Model dari data latih mendapatkan probabilitas sebagai berikut untuk masing-masing contoh dalam kedua kelas yang ada:

- positif: $\{0.9, 0.4, 0.7, 0.8\}$
- negatif: $\{0.1, 0.7, 0.2, 0.3, 0.2, 0.5, 0.3, 0.6\}$

Gambarkan ROC curves dengan menggunakan nilai-nilai batas (*threshold*) berikut: 0.00, 0.25, 0.45, 0.65, 1.00! [5 poin]

Untuk membantu Anda, silakan gunakan formula berikut:

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

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