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

Jurusan	:	Teknik Informatika	**Hari/Tanggal**	:	Kamis, 27 April 2017
Mata Kuliah	:	Data Mining	**Sifat**	:	Open book, kalkulator
Nama Dosen	:	Ir. Endang R., 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 (.)

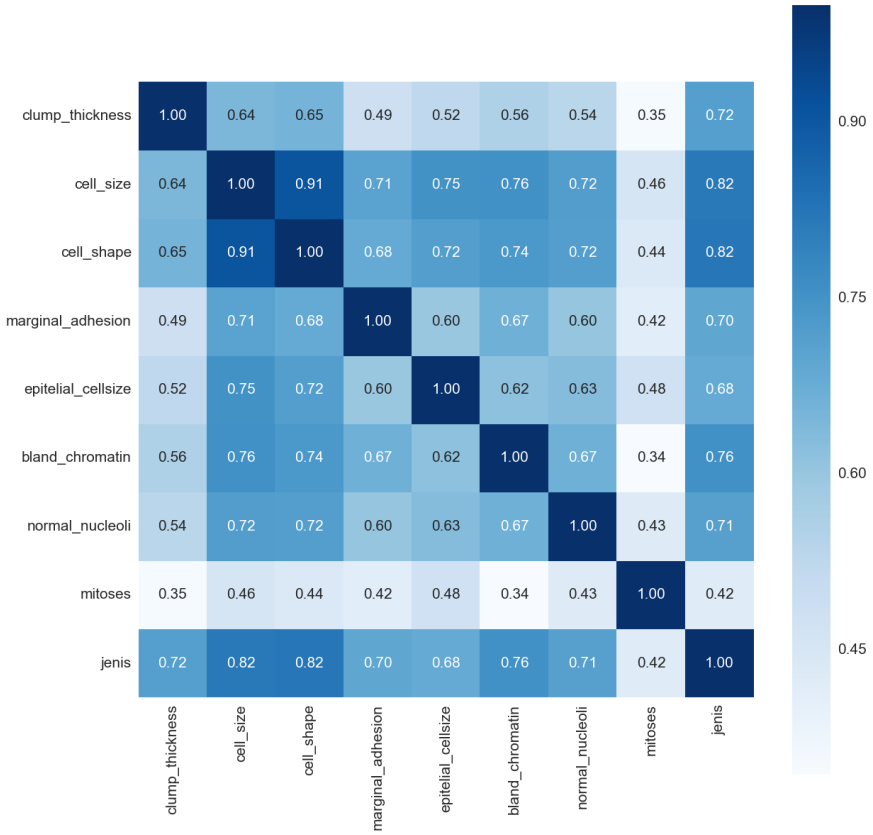
1 Analisis dan Representasi Data

Anda diberikan data tentang beberapa atribut yang menentukan apakah seseorang mengidap kanker payudara atau tidak. Dataset ini dikenal dengan nama Wisconsin Breast Cancer (WBC) dataset. Ada sembilan atribut dalam dataset tersebut, yaitu: *clump thickness*, *cell size*, *cell shape*, *marginal adhesion*, *epitelial cellsize*, *bare nuclei*, *bland chromatin*, *normal nucleoli*, dan *mitoses*. Kesembilan atribut tersebut digunakan untuk memprediksi dua jenis tumor: jinak (*benign*) atau ganas (*malignant*; dikenal juga dengan nama kanker).

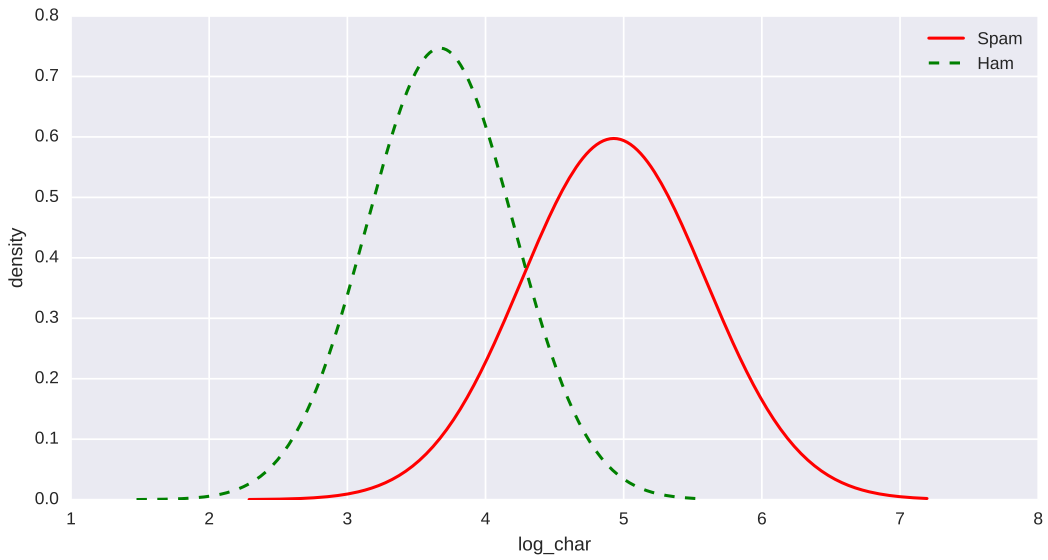
(a) Diberikan heatmap korelasi seperti pada Gambar 1. Apa yang dapat Anda ceritakan berdasarkan heatmap tersebut? [4 poin]

(b) Dari total 699 data, terdapat 10 data yang nilai *bare nuclei*-nya tidak diketahui. Apa yang akan Anda lakukan terhadap 10 data ini? Sebutkan semua alternatif yang dapat dilakukan! [4 poin]

(c) Kesembilan atribut dalam dataset tersebut memiliki nilai 1-10. Jika Anda ingin membuat *Nearest Neighbour classifier*, kira-kira konsep kedekatan data apa yang akan Anda gunakan? *Similarity* atau *distance*? Jenis yang mana? Berikan alasannya. [2 poin]



Gambar 1: Heatmap korelasi dari pasangan-pasangan atribut



Gambar 2: *Class-conditional* Gaussian dari logaritma jumlah karakter teks komentar

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 `has_pattern_yu+k` merupakan kecocokan *regular expression* (regex) "yu+k" dalam teks komentar dan `has_bbm_pin` adalah keberadaan PIN BlackBerry® Messenger dalam teks komentar. Asumsikan bahwa `has_bbm_pin` dan `has_pattern_yu+k` mengikuti distribusi Bernoulli. *Probability mass function* (PMF) dari distribusi Bernoulli adalah:

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

Tabel 1: Data latih filter komentar spam

has_bbm_pin	has_pattern_yu+k	label
True	True	spam
False	False	spam
True	False	spam
True	True	spam
False	False	spam
False	False	ham
False	False	ham
True	False	ham
False	False	ham
False	False	ham

(a) Hitunglah semua parameter distribusi Bernoulli yang diperlukan! Tunjukkan perhitungan Anda! [4 poin]

(b) Menggunakan model yang telah terbentuk di bagian (a), klasifikasikan teks komentar "Add PIN BBM gw di 7D48707E yuk"! Tunjukkan perhitungan Anda! [4 poin]

(c) Anda diberikan grafik dua distribusi Gaussian seperti pada Gambar 2 dengan sumbu x adalah fitur `log_char` yang berisi logaritma dari jumlah karakter teks komentar. Apa yang dapat Anda simpulkan dari grafik tersebut? [2 poin]

3 Decision Trees dan ROC Curves

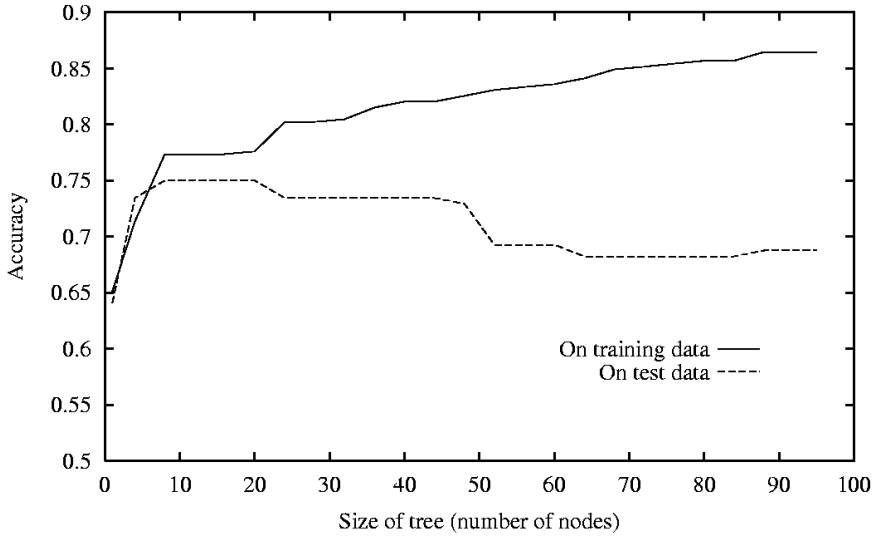
3.1 Decision Trees

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

(b) Perhatikan Gambar 3. Berapa ukuran pohon (*size of tree*) yang sebaiknya digunakan? Jelaskan! [2 poin]



Gambar 3: Kurva hubungan *size of tree* dengan akurasi (Mitchell, 1997)

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! [4 poin]