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

Program Studi	:	Teknik Informatika	**Hari/Tanggal**	:	Rabu, 8 November 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 diminta untuk menganalisis kemungkinan seorang pengguna layanan telekomunikasi untuk *churn*. *Churn* didefinisikan sebagai berhentinya seseorang dalam penggunaan layanan. Hal ini dapat terjadi karena tidak menggunakan layanan apapun sama sekali maupun karena pindah ke penyedia layanan yang lain.

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

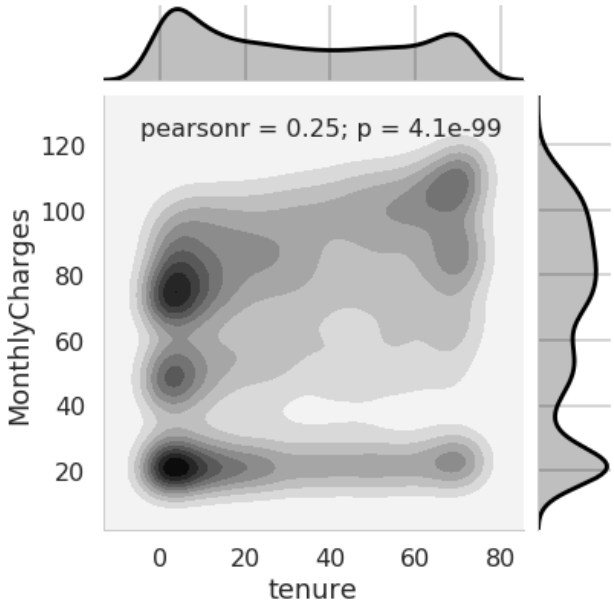
(b) Selain *tenure* (lama berlangganan dalam satuan bulan) dan *monthly charges*, Anda juga mempunyai kolom *total charges* di dalam tabel. Apakah Anda akan menggunakan nilai ini bersamaan dengan *tenure* dan *monthly charges* untuk memprediksi kemungkinan *churn*? Mengapa? [2 poin]

(c) Dalam data yang Anda punya, ditemukan bahwa terdapat cukup banyak atribut yang bersifat nominal (kategori). Berdasarkan pengetahuan tersebut, jika Anda ingin membuat *classifier* yang memberikan bobot pada tiap atribut, apa yang harus Anda lakukan terlebih dahulu dalam prapemrosesan? Sebagai contoh, jika Anda diberikan atribut *payment method* dengan

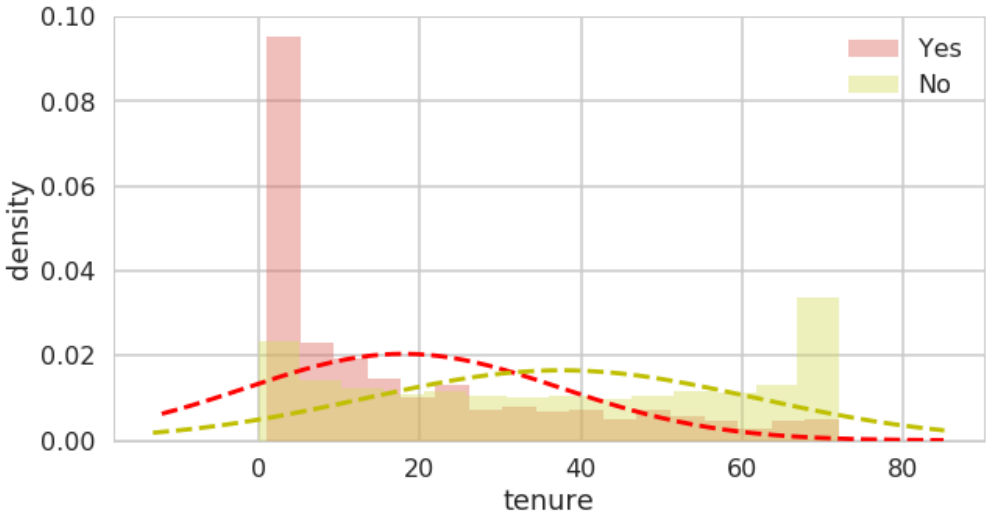
$$\text{payment method} \in \{\text{electronic check, mailed check, bank transfer, credit card}\},$$

seperti apa bentuk keluaran dari prapemrosesan yang Anda lakukan? [2 poin]

(d) Jika Anda diberikan kesempatan untuk mengumpulkan variabel baru untuk membantu analisis Anda, variabel apa yang ingin Anda cari dari pengguna? Mengapa? [2 poin]



Gambar 1: *Contour plot* dari *tenure* (dalam satuan bulan) dan *monthly charges*



Gambar 2: *Class-conditional* Gaussians dari *tenure*

2 Naïve Bayes & Decision Trees

Berdasarkan data yang diberikan pada soal 1, Anda membuat sebuah *classifier* untuk menghitung probabilitas seorang pengguna akan *churn* dengan menggunakan metode Naïve Bayes. Diberikan data seperti pada Tabel 1 dengan `tenure` menggambarkan lamanya pengguna tersebut telah berlangganan (dalam satuan bulan) dan `phone_service` adalah status apakah pengguna tersebut menggunakan layanan telepon. Asumsikan bahwa `tenure` terdistribusi normal dan `phone_service` 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/normal 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 prediksi churn

tenure	phone_service	churn
1	Yes	Yes
1	Yes	Yes
16	No	Yes
1	Yes	Yes
5	Yes	Yes
65	Yes	No
38	Yes	No
36	Yes	No
72	Yes	No
1	Yes	No

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

(b) Menggunakan model yang telah terbentuk di bagian (a), klasifikasikan data baru dengan `tenure=70` dan `phone_service=No`! Tunjukkan perhitungan Anda! [4 poin]

(c) Anda diberikan grafik dua distribusi Gaussian seperti pada Gambar 2 dengan sumbu x adalah fitur `tenure`. Apa yang dapat Anda simpulkan dari grafik tersebut? [2 poin]

(d) Menggunakan intuisi Anda, buatlah pohon keputusan yang dapat menghasilkan akurasi hingga 90% pada data latih yang diberikan. Apa yang menjadi alasan Anda dalam menghasilkan pohon keputusan tersebut? Apakah mungkin untuk menghasilkan akurasi hingga 100% dalam data latih ini? [4 poin]

(e) Apakah model yang Anda hasilkan pada bagian (d) juga akan mencapai akurasi yang sama secara umum untuk data yang baru? Berikan penjelasan Anda. [2 poin]

(f) Asumsikan bahwa Anda punya lebih banyak atribut daripada yang tertera pada Tabel 1. Seandainya kita melakukan *one-hot-encoding* terhadap semua atribut nominal yang ada dalam data, apakah kita masih perlu menggunakan *gain ratio*? Mengapa? [4 poin]

3 Evaluasi Model

Anda diminta untuk membuat sebuah sistem temu balik informasi (*information retrieval*) seperti Google. Jadi, pengguna akan memasukkan *query*, lalu model Anda akan mengembalikan sekumpulan dokumen yang dianggap relevan terhadap *query* tersebut. Umumnya, Anda akan menggunakan *precision* dan *recall* sebagai metrik untuk mengevaluasi model Anda. Akurasi dirumuskan sebagai

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

sedangkan *precision* dan *recall* dirumuskan sebagai

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

(a) Mengapa kita tidak menggunakan akurasi dalam kasus ini sebagai metrik evaluasi? [2 poin]

(b) Bagaimana kita dapat menginterpretasi nilai *precision* dan *recall* dalam kasus ini? [4 poin]

(c) Mengapa menjadi penting bagi kita untuk melaporkan *precision* dan *recall* sekaligus dan bukan hanya salah satu? [2 poin]

(d) Di kasus seperti apa lagi kita akan menggunakan metrik evaluasi seperti ini? [2 poin]