Review Skripsi 1

Topik Skripsi	: MTA4801 - Penerapan Algoritma Anonimisasi Pada Lingkungan Big Data
Nama Mahasiswa	: Stephen Jordan
NPM	: 2016730018
Pembimbing Utama	: Mariskha Tri Adithia, P.D.Eng
Pembimbing Pendamping	: Dr. Veronica Sri Moertini
Reviewer	: Husnul Hakim, M.T.

Presentasi

Secara umum, presentasi sudah cukup baik. Berikut ini saran-saran yang dapat dipertimbangkan untuk membuat presentasi menjadi lebih baik:

1. Penjelasan latar belakang dan tujuan tidak jelas padahal penjelasan di dokumen kemajuan skripsi sudah ditulis dengan cukup baik.
2. Penjelasan tahapan Naïve Bayes dan k-Means tidak jelas karena tidak terdapat poin-poin yang menyatakan tahapan dari kedua teknik tersebut.
3. Mungkin perlu ditambahkan juga hasil analisis diagram kelas yang sudah dibuat.

Dokumen

Tata Tulis Laporan

Secara umum, tata tulis laporan sudah baik. Adapun catatan-catatan penting mengenai dokumen adalah sebagai berikut:

1. Beberapa kalimat masih terlalu panjang. Kalimat-kalimat seperti ini dapat diperbaiki dengan memecahnya menjadi beberapa kalimat yang lebih sederhana. Perhatikan komentar di dokumen sripsi yang sudah di-*review* untuk lebih jelasnya!
2. Masih terdapat beberapa kesalahan ketik, misalnya pada latar belakang ada kata dPenalaran, ada kata membandingkat pada penjelasan tentang k-Means, dan lain-lain.
3. Perhatikan cara penulisan imbuhan asing: antar, sub, anti, dan lain-lain. Pada dokumen yang telah di-*review* telah diberikan tautan sebagai panduan penulisan.
4. Masih ada tabel-tabel yang dibuat sebagai gambar.
5. Penulisan notasi matematika juga masih ada yang berupa tangkapan layar.
6. Setiap persamaan atau rumus-rumus harus diberi nomor rumus atau nomor persamaan. Contoh persamaan yang belum diberi nomor adalah pada langkah (b) halaman 9.
7. Ada persamaan yang dituliskan berulang-ulang, contohnya pada langkah (b) dan (c) pada halaman 10. Hal ini dapat dihindari dengan memberi nomor persamaan.
8. Ada urutan yang penjelasan yang harus diatur ulang. Misalnya tentang information lost. Istilah dan rumus untuk menghitung information loss sudah ada di halaman 10, namun penjelasannya baru ada di halaman 13.

Konten

Berikut ini adalah perbaikan yang perlu dilakukan untuk konten dokumen kemajuan skripsi:

1. Langkah-langkah dari Naïve Bayes harus diperbaiki! Langkah-langkah harus berupa proses atau tahapan. Tidak ada tahapan yang merupakan gambar. Misalnya pada langkah b, tertulis: "(b) Gambar 3 adalah contoh menghitung probabilitas masing-masing atribut."
2. Pada Gambar 3 yang terdapat pada halaman 5, sepertinya terdapat kesalahan perhitungan frekuensi dan probabilitas, jika perhitungan didasarkan pada data pada Gambar 2 yang terdapat pada halaman 4.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes

Gambar 2: Dataset Kondisi Cuaca Bermain Golf

Jika data tersebut yang digunakan, sepertinya seharusnya perhitungan frekuensi dan probabilitas adalah sebagai berikut:

Outlook

	Yes	No	P(Yes)	P(No)
Sunny	2	1	$\frac{2}{5}$	$\frac{1}{4}$
Overcast	2	0	$\frac{2}{5}$	0
Rainy	1	3	$\frac{1}{5}$	$\frac{3}{4}$
Total	**5**	**4**	100%	100%

Sementara yang terdapat di dokumen skripsi adalah sebagai berikut:

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Pada tabel Outlook yang ada di dokumen skripsi, tertulis: ada 3 outlook sunny dengan play golf no, padahal hanya 1, yaitu data nomor 5 dari Gambar 2.

Jika dilihat, sepertinya data yang diambil dari website Geeks For Geeks. Ada sebanyak 14 buah baris data pada website tersebut, namun yang ditampilkan di dokumen hanya 9 buah baris data. Tetapi, pada dokumen skripsi, untuk contoh perhitungan, digunakan hasil perhitungan yang juga terdapat di website, yaitu perhitungan frekuensi dan probabilitas untuk 14 buah baris data. Sayangnya, apabila diihat lebih teliti hasil perhitungan yang disajikan pada website juga masih ada kesalahan.

Berdasarkan hal itu, disarankan agar penjelasan pada dasar teori tidak diambil dari halaman web, namun dari buku teks.

3. Langkah-langkah dari k-Means diperbaiki! Langkah-langkah harus berupa proses atau tahapan. Tidak ada tahapan yang merupakan gambar. Misalnya pada langkah a, tertulis: " Gambar 5 adalah hasil pengelompokan awal untuk k = 2." Perlu juga dijelaskan tentang kondisi berhentinya proses *clustering* (saat banyaknya iterasi sudah mencapai maksimum, dan saat sudah tidak ada perbedaan hasil *clustering* pada iterasi ke-i dan ke-$(i-1)$.
4. Langkah pertama dari k-Means tidak jelas. Tertulis: " Untuk menentukan titik centroid awal, akan dicari nilai A dan B terjauh dengan data lainnya menggunakan Euclidean distance. Bagaimana cara mendapatkan nilai A dan B yang terjauh dengan data lainnya? Apa yang dimaksud dengan 2 data yang terjauh dengan data lainnya?
5. Langkah (c) dari k-Means tidak jelas. Tertulis: " Menentukan titik centroid baru pada cluster yang baru terbentuk dari tahap sebelumnya." Harus dijelaskan bagaimana caranya menentukan titik centroid baru. Apakah dengan menghitung rata-rata dari tiap member pada masing-masing *cluster*? Apakah sama seperti langkah (a)?
6. Sebelum masuk ke penjelasan tentang Greedy k-Member Clustering, harus dijelaskan terlebih dahulu tentang anonimisasi data, beserta istilah-istilah yang ada di dalamnya, seperti PII, quasy identifier, dan lain-lain.
7. Teorema 1 dan bukti pada penjelasan tentang Greedy k-Member Clustering tidak jelas hubungannya.
8. Masing-masing tipe data pada Scala yang disebutkan pada halaman 24 perlu dijelaskan satu persatu.
9. Contoh "Membuat Kelas Object pada Scala" di halaman 25 perlu diperbaiki. Contoh yang diberikan adalah pembuatan *singelton object*. Jika yang dimaksud adalah pembuatan kelas, seharusnya contoh yang diberikan tidak seperti itu.
10. Pada halaman 27, perlu dijelaskan:
    a. data apa yang akan dikelompokkan dengan Naïve Bayes
    b. banyak datanya ada berapa,
    c. berapa banyak data yang dijadikan *data test*
    d. berapa banyak yang menjadi *data training*

    sehingga hasil akhir yang ditampilkan di dokumen sebagai Gambar 26 dapat dipahami dengan lebih jelas.

11. Pada halaman 28, perlu dijelaskan:
    a. data apa yang akan dikelompokkan dengan k-Means
    b. banyak datanya ada berapa,

c. berapa banyak maksimum iterasi yang digunakan
d. berapa banyak clusternya

sehingga hasil akhir yang ditampilkan di dokumen sebagai Gambar 27 dapat dipahami dengan lebih jelas.

Pertanyaan

1. Pada halaman 5 terdapat pernyataan: "teorema Bayes saling independen terhadap fitur-fiturnya". Apa yang dimaksud dengan saling independen?

   Jawaban:
   Teorema Bayes pada dasarnya menghitung probabilitas bersyarat untuk masing-masing fitur. Dalam menghitung probabilitas bersyarat, nilai yang dibutuhkan adalah peluang kemunculan fitur dengan nilai label “yes” atau “no” saja dan tidak bergantung pada peluang kemunculan nilai fitur lainnya sehingga teorema bayes bersifat saling independen terhadap fitur-fiturnya.

2. Berapakah nilai k terbaik untuk melakukan *clustering* dengan k-Means?

   Jawaban:
   NIlai k terbaik harus ditentukan melalui beberapa eksperimen berdasarkan hasil pengelompokan yang diinginkan. Pada kasus ini, hasil pengelompokan yang diinginkan adalah data-data yang memiliki nilai label yang sama, harus menjadi satu kelompok.

3. Terdapat 1.000.000 data yang akan di-*cluster* dengan menggunakan k-Means. Jika dilakukan *clustering* berkali-kali pada data tersebut dengan menggunakan k-Means, apakah *cluster* yang diperoleh akan selalu sama? Yang dimaksud dengan clustering berkali-bukan iterasinya, tapi dilakukan k-Means clustering berkali-kali. Misalnya 7 kali dengan banyak iterasi masing-masing 1000. Apakah hasil dari 7 kali clustering itu akan selalu sama?

   Jawaban:
   Apabila dilakukan 7 kali clustering dengan masing-masing iterasi berjumlah 1000 kali (dengan catatan nilai k sama untuk masing eksperimen), maka hasil akhir pengelompokan akan selalu sama. Nilai k dapat menentukan hasil akhir dari pengelompokan clustering karena membatasi jumlah data pada sebuah kelompok data yang mempengaruhi hasil pencarian masing-masing data yang dekat dengan centroid dari kelompok data tertentu.

4. Sebuah algoritma greedy memang dapat memiliki kompleksitas yang lebih baik dari teknik lainnya, namun biasanya belum tentu menghasilkan solusi yang optimal. Apakah Algoritma Greedy k-Member Clustering menghasilkan solusi optimal?

   Jawaban:
   Optimal dalam kasus ini adalah mendapatkan nilai informasi yang tinggi. Algoritma Greedy k-member clustering dapat menghasilkan hasil yang optimal karena membentuk masing-masing kelompok data berdasarkan nilai information loss paling rendah. Pada progress report saya

mencantumkan bahwa algoritma Greedy k-Member Clustering belum memiliki hasil yang optimal karena masih ada algoritma pengelompokan data lain yang dapat memberikan nilai informasi yang lebih tinggi dari pengelompokan data pada Greedy k-Member Clustering

5. Apa yang dimaksud dengan NP-Hard?

Jawaban:
NP-Hard adalah komplesitas algoritma Greedy k-member clustering dapat mencapai kompleksitas yang eksponensial untuk mencari pengelompokan data yang baik (berdasarkan nilai information loss dan distance) sehingga apabila tidak dijalankan pada sistem terdistribusi yang berjalan secara paralel, makan komputasinya akan berjalan sangat lambat.

6. Sebutkan contoh kasus penggunaan *immutable variable* pada Scala!

Jawaban:
Scala berhubungan dengan operasi pada Spark. Contoh kasus immutable variable pada Scala hanya digunakan apabila ingin menyimpan hasil operasi transformasi pada Spark, contohnya adalah fungsi map() yang dapat memetakan nilai key, value. Fungsi tranformasi memiliki hasil yang selalu tetap dan tidak akan berubah sehingga immutable variable cocok untuk menyimpan hasil dari transformasi.

7. Dalam konsep pemrograman berorientasi objek, apa bedanya objek dengan kelas?

Jawaban:
Sebenarnya objek dan kelas prinsipnya sama saja (tIdak ada perbedaan) pada pemrograman berorientasi objek yaitu memiliki method dan inner variable (private int ….) pada kelas maupun objek. Bedanya adalah saat digunakan pada pemrograman Spark. Penggunaan objek lebih sering digunakan untuk melakukan pemrograman Spark terutama saat saya praktikum mata kuliah Analisis Big Data karena lebih umum untuk digunakan. Penggunaan objek dan kelas tidak mempengaruhi hasil pada pemrograman Spark, sehingga keduanya dapat digunakan.