Hafizh Renanto Akhmad
13621060

AE2204 SAINS DATA DAN STATISTIKA
TUGAS MANDIRI

0. Pertanyaan

   Jawaban

A. Pemelajaran Statistika

1. Jelaskan dengan contoh, apa yang dimaksud pemelajaran statistika.

   Pemelajaran statistika atau *statistical learning* adalah kumpulan metode dan algoritma yang bertujuan untuk mengestimasi sebuah relasi/fungsi antara sekumpulan input dan sekumpulan output berdasarkan data-data yang ada.

   Misalkan kita memiliki sekumpulan input $X = (X_1, X_2, \dots, X_p)$ dan sekumpulan output $Y = (Y_1, Y_2, \dots, Y_p)$. Terdapat sebuah fungsi f yang memetakan X ke Y (menggambarkan relasi antara X dan Y), atau

   $$Y = f(X) + \epsilon \tag{1}$$

   dimana ϵ adalah *error*. Pemelajaran statistik adalah kumpulan metode yang digunakan untuk mengestimasi fungsi f.

   Contoh: Pada proses desain bangunan di Indonesia sebagai negara yang dilalui Ring of Fire, perlu diperhitungkan factor ketahanan bangunan terhadap gempa. Oleh karena itu, dikumpulkan data mengenai bangunan-bangunan yang sebelumnya pernah terkena gempa. Data-data yang dimaksud adalah data-data kondisi sebelum dan setelah gempa terjadi, diantaranya: kondisi lantai sebelum gempa, tinggi bangunan sebelum gempa, dan material yang digunakan dalam membangun bangunan, serta tingkat kerusakan yang terjadi setelah gempa terjadi. Data-data tersebut akan digunakan untuk membantu memperoleh hubungan antara tingkat kerusakan dengan parameter-parameter lain.

2. Dari suatu himpunan data, mengapa kita perlu/ ingin mengestimasi relasi antara masukan dan respon? Secara umum, bagaimana cara mengestimasi relasi antara masukan dan respon? Jelaskan jawaban Anda.

Ada dua alasan utama mengapa kita mengestimasi relasi antara masukan dan respon, yakni untuk *prediksi* dan untuk *inferensi*.

Berikut ilustrasi situasi ketika estimasi relasi input dan output diperlukan:

- Misalkan kita memiliki beberapa data input X yang tersedia dan ingin mengetahui nilai output Y yang bersesuaian. Namun, nilai output Y tersebut tidak bisa didapatkan dengan mudah. Dengan demikian, nilai Y tersebut dapat diprediksi menggunakan (1),

$$\hat{Y} \approx \hat{f}(X) \tag{2}$$

  dengan $\hat{f}$ adalah estimasi dari f dan $\hat{Y}$ adalah hasil prediksi dari Y. Pada kasus ini, kita mengestimasi f untuk memprediksi nilai output.
- Pada kasus lain, kita hanya ingin memahami bagaimana Y dipengaruhi oleh X. Dengan demikian, f perlu diestimasi, tapi tidak harus memperoleh bentuk seperti (2). Pada kasus ini, kita mengestimasi f untuk memahami bagaimana hubungan nilai-nilai input dengan nilai output.

f bisa diestimasi dengan menggunakan beberapa data points yang sudah ada dan mengaplikasikan pemlejaran statistic menggunakan data-data tersebut. Ada dua pendekatan dalam melakukan hal tersebut:

a. Parametric. Pada pendekatan ini, metode yang digunakan mengasumsikan bentuk fungsi f. Salah satu contohnya adalah mengasumsikan hubungan linear. Setelah menentukan asumsi bentuk, bisa dicari parameter-parameter yang muncul akibat asumsi tersebut.
b. Non-parametric. Pada pendekatan ini, metode yang digunakan tidak mengasumsikan bentuk fungsi f, melainkan mengestimasi fungsi f yang paling mewakili data points sembari menghindari terlalu kasar.

3. Untuk tiap bagian a) sampai c), tunjukkan secara umum apakah kita mengharapkan performa dari metode pemelajaran statistika fleksibel untuk lebih baik atau buruk dibandingkan dengan metode yang tidak fleksibel. Berikan alasan atas jawaban Anda.

   a) Ukuran sampel n sangat besar, dan jumlah variabel desain p kecil.

   Pada kasus ini, metode pemelajaran fleksibel diharapkan memiliki performa lebih baik dari metode tidak fleksibel. Hal ini karena, dengan ukuran sampel n yang sangat besar, ada kemungkinan terdapat banyak bentuk-bentuk yang tidak bisa ditangkap oleh asumsi dari metode tidak fleksibel.

   b) Jumlah variabel desain p sangat besar, dan jumlah observasi n kecil.

Pada kasus ini, metode pemelajran tidak fleksibel diharapkan memiliki performa lebih baik dari metode tidak fleksibel. Hal ini karena dengan jumlah n yang kecil namun p sangat besar, metode fleksibel kemungkinan besar akan mengalami *overfitting* data, membuat performanya menurun ketika digunakan untuk berbagai data dengan distribusi yang bermacam-macam.

c) Hubungan antara variabel desain dan respon sangat nonlinear.

Pada kasus ini, metode pemelajaran fleksibel diharapkan memiliki performa lebih baik dari metode tidak fleksibel. Hal ini karena pemelajaran tidak fleksibel memiliki asumsi respons yang terbatas, dan mungkin tidak menangkap hubungan kompleks yang sangat nonlinear secara akurat.

4. Sekarang pikirkan beberapa aplikasi riil dari pemelajaran statistika di bidang kedirgantaraan.

a) Deskripsikan tiga aplikasi nyata di bidang kedirgantaraan di mana klasifikasi akan berguna. Deksripsikan respon dan juga variabel desainnya. Apakah tujuan dari tiap aplikasi inference atau prediksi? Terangkan jawaban Anda.

- Failure Prediction. Pada kasus ini, pemelajaran statistika digunakan untuk memprediksi apakah suatu alat dalam pesawat sedang/akan mengalami kerusakan, dengan menggunakan variabel design seperti data-data dari sensor (tekanan, suhu, dst.) dan riwayat *maintenance* yang pernah dilakukan. Proses ini dapat membantu maintenance di industry kedirgantaran. Inferensi juga bisa dilakukan, seperti menghungkan parameter dari sensor tertentu dengan kerusakan yang dialami.
- *Computer Vision* untuk Deteksi Bagasi di Bandara. Variabel desain adalah gambar-gambar (yang merupakan matriks dengan nilai-nilai terentu, seperti kode warna RGB) yang terdapat bagasi dan tidak terdapat bagasi di dalamnya, dan variabel responsnya adalah penentuan apakah pada gambar yang ditangkap oleh kamera terdapat bagasi atau tidak. Tujuan dari aplikasi ini adalah prediksi.
- Prediksi Delay Penerbangan Pesawat. Variabel desainnya antara lain, tanggal berangkatan, jam keberangkatan, jarak tempuh keberangkatan, dan jenis maskapai. Variabel responsnya adalah penentuan apakah penerbangan tersebut akan delay atau tidak. Tujuan dari aplikasi: prediksk

b) Deskripsikan tiga aplikasi nyata di bidang kedirgantaraan di mana regresi akan berguna. Deksripsikan respon dan juga variabel desainnya. Apakah tujuan dari tiap aplikasi inference atau prediksi? Terangkan jawaban Anda.

- Desain Sayap Pesawat Terbang. Dalam desain sayap suatu pesawat, nilai-nilai koefisien aerodinamik (C_L, C_D, C_M) sangat penting untuk diketahui. Variabel desain yang digunakan

adalah bentuk geometri dari airfoil (e.g. radius leading edge, upper crest curvature, dst.), dan variabel responnya adalah koefisien aerodinamik yang telah disebutkan. Tujuan dari aplikasi ini adalah untuk memprediksi nilai koefisien aerodinamik tersebut.

- Prediksi Kondisi Cuaca. Pada kasus variabel desain yang digunakan adalah data cuaca seperti kecepatan angin, kelembapan udara, kelembapan tanah, dan Riwayat curah hujan hari-hari sebelumnya. Output yang dihasilkan adalah curah hujan. Tujuan dari aplikasi ini adalah untuk memprediksi curah hujan disuatu hari yang bisa membantu mengambil keputusan yang berhubungan dengan penerbangan, seperti dalam menentukan jadwal dan rute penerbangan.
- Prediksi Thrust Sistem Propulsi Pesawat. Variabel desainnya adalah parameter-parameter yang dimiliki komponen-komponen pada sistem propulsi. Variabel respons/outputnya adalah nilai thrust yang dihasilkan. Tujuan dari aplikasi ini adalah prediktif. Aplikasi ini digunakan dalam desain system propulsi sehingga dicapai tingkat *safety* dan *reliability* yang tinggi.

B. Regresi Linear

5. Misalkan kita mempunyai suatu himpunan data dengan lima variabel desain, X_1 = GPA, X_2 = IQ, X_3 = Level (1 untuk College and 0 untuk High School), X_4 = Interaksi antara GPA and IQ, and X_5 = Interaksi antara GPA and Level. Responnya adalah gaji awal setelah kelulusan (dalam ribu dolar). Misalkan kita menggunakan *least squares* untuk memprediksi model, dan mendapatkan $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, $\beta_5 = -10$.
   a. Mana jawaban yang benar, dan mengapa?
      i Untuk IQ dan GPA bernilai tetap, rata-rata lulusan *high school* memperoleh gaji lebih dari lulusan *college*.
      ii Untuk IQ dan GPA bernilai tetap, rata-rata lulusan *college* memperoleh gaji lebih dari lulusan *high school*.
      iii Untuk IQ dan GPA bernilai tetap, rata-rata lulusan *high school* memperoleh gaji lebih dari lulusan *college* jika GPA-nya cukup tinggi.
      iv Untuk IQ dan GPA bernilai tetap, rata-rata lulusan *college* memperoleh gaji lebih dari lulusan *high school* jika GPA-nya cukup tinggi.

      Misalkan

      x_1 adalah nilai GPA,

      x_2 adalah nilai IQ,

x_3 adalah nilai level pendididkan (0 untuk *high school* dan 1 untuk *college*),

x_4 adalah nilai interaksi antara GPA dan IQ, yakni, $x_4 = x_1 \cdot x_2$, dan

x_5 adalah nilai interaksi antara GPA dan level, yakni $x_5 = x_1 \cdot x_3$,

untuk suatu *data points*. Maka, nilai gaji awal setelah kelulusan (y) bisa diestimasi dengan

$$\begin{aligned}\hat{y} = \widehat{\text{gaji}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ &= 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5 \\ &= 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_1x_2 - 10x_1x_3\end{aligned}$$

Jika GPA dan IQ dibuat tetap dengan masing-masing nilai a dan b,

$$\hat{y}(X_1 = a, X_2 = b) = 50 + 20a + 0.07b + 35x_3 + 0.01ab - 10ax_3$$

Untuk lulusan *college*,

$$\hat{y}(X_1 = a, X_2 = b, X_3 = 1) = \hat{y}_{\text{college}} = 85 + 20a + 0.07b + 0.01ab - 10a$$

Untuk lulusan *high school*,

$$\hat{y}(X_1 = a, X_2 = b, X_3 = 1) = \hat{y}_{\text{high school}} = 50 + 20a + 0.07b + 0.01ab$$

Selisih antara keduanya yakni,

$$\hat{y}_{\text{college}} - \hat{y}_{\text{high school}} = 35 - 10a$$

Terlihat bahwa selisih estimasi gaji untuk lulusan *college* dan lulusan *high school*, jika IQ dan GPA dibuat tetap, bergantung pada nilai GPA. $\hat{y}_{\text{college}}$ akan lebih kecil jika

$$\begin{aligned}\hat{y}_{\text{college}} - \hat{y}_{\text{high school}} &< 0 \\ 35 - 10a &< 0 \\ 10a &> 35 \\ a &> 3.5\end{aligned}$$

Dengan demikian, gaji lulusan *high school* bisa lebih besar daripada gaji lulusan *college* untuk nilai IQ dan GPA yang tetap, jika lulusan tersebut memiliki GPA yang tinggi, tepatnya lebih dari 3.5 (jawaban iii.)

b. Prediksi gaji dari seorang lulusan college dengan IQ 110 dan GPA 4,0.

Dengan menggunakan persamaan sebelumnya,

$$\begin{aligned}\widehat{\text{gaji}}(X_1 = 4, x_2 = 110, X_3 = 1) &= 85 + 20(4) + 0.07(110) + 0.01(4)(110) - 10(4) \\ &= 137.1\end{aligned}$$

Diprediksi bahwa gaji dari seorang lulusan college dengan IQ 110 dan GPA 4,0 adalah $137.1.

c. Benar atau salah: karena koefisien dari interaksi antara GPA dan IQ sangat kecil, terdapat bukti yang sangat sedikit tentang adanya interaksi. Berikan alasan atas jawaban anda.

Salah, karena koefisien tersebut hanya menandakan pengaruh interaksi antara GPA dan IQ terhadap gaji awal, bukan hubungan interaksi antara keduanya.

6. Jelaskan dengan baik perbedaan dari metode KNN untuk klasifikasi dan KNN untuk regresi.

   KNN adalah salah satu pemelajaran statistik yang cukup luas dipakai. KNN bekerja dengan menghitung sebuah parameter jarak antara titik yang akan diprediksi dengan titik-titik yang ada. Hasil prediksi bergantung pada hasil "vote" tertinggi dari k titik-titik terdekat yang ada (titik tetangga). Untuk klasifikasi, "vote" tertinggi yang dimaksud adalah kelas dengan frekuensi muncul tertinggi dari k titik tetangga. Untuk regresi, "vote" yang dimaksud adalah rata-rata dari nilai-nilai yang dimiliki oleh k titik tetangga.

C. Klasifikasi

7. Misalkan kita mengumpulkan data sebuah kelompok di kelas statistika dengan variabel X_1 = jam belajar, X_2 = GPA kuliah sarjana, and Y = memperoleh nilai A. Kita membuat regresi logistik $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.

   Distribusi probabilitas untuk regresi logistik adalah:

$$p(X_1 = x_1, \dots, X_p = x_p) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p}}$$

   dengan $X_1, \dots, X_p$ adalah *random variable* tiap input.

   a. Perkirakan berapa probabilitas bahwa seorang siswa yang belajar 40 jam dan mempunyai GPA kuliah sarjana 3,5 mendapatkan nilai A di kelas tersebut.

$$p(X_1 = 40, X_2 = 3.5) = \frac{1}{1 + e^{6 - 0.05 \cdot 40 - 1 \cdot 3.5}} = 0.3775$$

   Kemungkinan seorang siswa yang belajar 40 jam dan mempunyai GPA kuliah sarjana 3,5 mendapatkan nilai A adalah 0,3775.

   b. Berapa banyak jam belajar yang diperlukan siswa di bagian (a) untuk mendapatkan 50% kemungkinan mendapatkan nilai A di kelas tersebut?

$$\begin{aligned} p(X_1 = \alpha, X_2 = 3.5) &= 0.5 \\ \frac{1}{1 + e^{6 - 0.05 \cdot \alpha - 1 \cdot 3.5}} &= 0.5 \\ 2 &= 1 + e^{6 - 0.05\alpha - 3.5} \\ e^{6 - 0.05\alpha - 3.5} &= 1 \\ 6 - 0.05\alpha - 3.5 &= 0 \end{aligned}$$

$$\alpha = 50$$

Jika seorang siswa memiliki GPA kuliah sarjana 3,5, maka banyak jam belajar yang diperlukan sehingga ia mendapatkan 50% kemungkinan untuk mendapat nilai A adalah 50 jam.

8. Soal ini berhubungan dengan *odds*.

   Odds suatu *event* didefinisikan sebagai rasio dari peluang suatu *event* terjadi dengan peluang suatu *event* tidak terjadi:

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)}$$

   dengan A adalah *event* yang dilihat, $P(A)$ adalah peluang A terjadi, dan $\text{odds}(A)$ adalah *odds* A terjadi.

   a. Rata-rata, berapa fraksi orang dengan *odds* gagal bayar atas tagihan kartu kredit sebesar 0,37 akan benar-benar gagal bayar?

      Pada kasus ini, $A =$ orang gagal bayar atas tagihan kartu kredit, dan $\text{odds}(A) = 0.37$, sehingga

$$0.37 = \frac{P(A)}{1 - P(A)} \Longrightarrow P(A) = \frac{0.37}{1 + 0.37} \approx 0.270073$$

      Maka secara rata-rata, fraksi orang dengan *odds* gagal bayar atas tagihan kartu kredit sebesar 0.37 yang akan benar-benar gagal bayar adalah 27%.

   b. Misalkan seseorang mempunyai kemungkinan 16% gagal bayar atas tagihan kartu kreditnya. Berapa *odds* bahwa orang tersebut akan gagal bayar?

      Pada kasus ini, $A =$ orang gagal bayar atas tagihan kartu kredit, dan $P(A) = 0.16$

$$\text{odds}(A) = \frac{0.16}{1 - 0.16} \approx 0.190476$$

      Maka *odds* orang dengan kemungkinan 16% gagal bayar atas tagihan kartu kreditnya adalah sekitar sebesar 0.19.