Open Recruitment DSAI

Prediksi Harga Laptop Berdasarkan Fitur-Fitur Tertentu.

Iron Man

Nugroho Adi Susanto (23/520312/PA/22367)

DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA

1. **Pertanyaan pertama: Dengan menggunakan kalimat Anda sendiri, jelaskan yang dimaksud dengan EDA (*Exploratory Data Analysis*)! Mengapa hal tersebut penting dalam proses analisis data?**

Explanatory data analysis adalah sebuah tindakan yang dilakukan oleh *data scientist* untuk menyelidiki dan menganalisis sebuah kumpulan data atau bisa disebut juga dengan *dataset.* Umumnya, EDA digunakan untuk menggali pola dan menemukan anomali dalam kumpulan data. Dalam dataset yang diberikan, EDA digunakan untuk melakukan analisis awal dalam menemukan korelasi antar *feature* yang ada dalam dataset. EDA menjadi langkah yang penting untuk menentukan perlakuan yang akan diberikan kepada masing-masing kolom dalam sebuah dataset. Dalam hal ini, penulis memperlakukan kolom berat (dengan mengekstraksi dan mengubah tipe data dalam kolom berat menjadi float) dan kolom harga sebagai sebuah kolom data yang bersifat kontinu. Penulis juga menemukan bahwa komponen-komponen lain dalam dataset dapat diperlakukan sebagai *categorical feature* karena memiliki sifat yang diskrit. Misalnya, penulis memperlakukan RAM sebagai sebuah tipe data diskrit. RAM memiliki kapasitas yang merupakan bilangan bulat kelipatan dua sehingga bisa diperlakukan sebagai data diskret. Hal-hal tersebut dapat diketahui dengan EDA.

2. Jelaskan perbedaan antara *supervised learning*, *unsupervised learning*, dan *reinforcement learning*! Termasuk kategori yang manakah *problemset* pada penugasan open recruitment ini?

Secara sederhana, *supervised learning* adalah metode pembelajaran mesin yang melatih model dengan dataset yang sudah dilabeli. Pada setiap baris data latihan, akan terdapat pasangan lengkap *feature* yang digunakan sebagai input dengan feature yang akan diprediksi. Setelah model dilatih, model akan diberikan feature input pada dataset input yang tidak pernah digunakan untuk melatih model dengan tujuan untuk memprediksi suatu feature pada dataset yang baru.

Tidak seperti *supervised learning*, *unsupervised learning* adalah metode pembelajaran mesin yang melatih model *machine learning* pada dataset yang tidak dilabeli. Oleh karena itu tidak akan ada *feature output* atau *feature* prediksi dalam dataset latihan. Sehingga model harus mempelajari pola *feature output* murni menggunakan *input feature*.

Reinforcement learning adalah metode pembelajaran mesin yang menempatkan suatu model dalam sebuah lingkungan *sandbox* yang umumnya berupa sebuah game dengan *reward function* yang sudah ditentukan. *Reward function* dibuat untuk mengarahkan model pada tindakan atau aksi yang diinginkan oleh developer. Pada awalnya, model tidak mengetahui apapun mengenai *reward function* tersebut dan bertindak sangat

acak. Namun, setelah melalui banyak percobaan, model akan belajar untuk memaksimalkan penghargaan yang Ia dapatkan melalui fungsi tersebut.

Menurut penulis, *problemset* pada penugasan ini termasuk dalam kategori supervised learning yang ditandai dengan adanya pasangan *feature input* dan *feature output* pada *training set* yang diberikan panitia.

3. **Apa yang dimaksud dengan *overfitting* dan *underfitting* dalam konteks *machine learning*? Apakah dalam pengerjaan penugasan praktek Anda mengalami salah satu atau kedua masalah tersebut? Bagaimana Anda menanganinya?**

Overfitting terjadi ketika model machine learning banyak terpengaruh oleh *noise* dalam suatu dataset hingga menurunkan akurasinya dalam memprediksi suatu data baru. Hal ini biasanya terjadi pada model dengan kompleksitas yang tinggi seperti *ensemble tree* dan *gradient boosting*, model dengan kompleksitas ini memiliki sensitifitas yang sangat tinggi sehingga membuatnya sensitif terhadap noise dalam data. Sensitifitas tersebut membuat akurasi model tersebut sangat mudah dipengaruhi oleh noise yang ada. *Overfitting* juga dapat terjadi karena minimnya data yang tersedia yang membuat model kesulitan menemukan pola yang cocok untuk data tersebut. *Overfitting* juga dapat terjadi karena adanya *feature* yang tidak relevan dengan *feature* prediksi dalam data.

Sementara itu, underfitting terjadi ketika model *machine learning* tidak mampu menangkap detail dalam training data. Hal ini umumnya terjadi karena model yang digunakan tidak cocok dengan dataset yang digunakan. Hal ini umumnya terjadi ketika data yang memiliki persebaran polynomial diproses menggunakan model regresi atau klasifikasi yang menggunakan fungsi berbasis linear.

Pada kasus *Ridge regression*, *overfitting* dan *underfitting* akan sangat berkaitan dengan koefesien *alpha* yang digunakan. Sebagaimana model regresi lain, model ini memiliki bentuk persamaan berupa

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \qquad (3.1)$$

Image 1.0 general linear regression equation in Elements of Statistical learning

Pada dasarnya akan dicari kombinasi dari variabel-variabel tersebut untuk meminimalisasi fungsi *error* yang didefinisikan dengan persamaan

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \Big)^2. \end{aligned} \qquad (3.2)$$

Image 2.0 Residual Sum of Squares in Elements of Statistical learning

Namun, pada model ini akan terdapat penalti yang diberlakukan pada setiap variabel untuk memanipulasi kompleksitas model. Penalti tersebut dinyatakan dalam.

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \qquad (3.52)$$

Image 3.0 Penalty function of ridge regressor in Elements of statistical learning.

Penulis sempat mengalami *overfitting* saat penulis menggunakan parameter alpha yang terlalu besar, mengingat alpha yang terlalu besar akan menyebabkan penurunan koefesien variabel pada model. Hal ini akan menyebabkan sejumlah model yang penting kehilangan pengaruhnya pada prediksi model. Beberapa variabel sangat berpengaruh terhadap harga menjadi tidak diperhitungkan lagi.

Sedangkan, *underfittng* penulis menggunakan nilai alpha yang terlalu rendah, hal ini menurunkan kompleksitas dan koefesien variabel yang menyebabkan sejumlah data yang tidak signifikan mempengaruhi model sehingga meningkatkan *root mean square error* pada evaluasi model.

Strategi yang dilakukan penulis untuk meminimalisasi *overfitting* dan *underfittng* adalah dengan melakukan gridsearch dengan kombinasi *alpha* = [1e-15, 1e-10, 1e-8, 1e-4, 1e-3,1e-2, 1, 5, 10, 20]. Setiap nilai *alpha* akan digunakan untuk melatih model yang akan diuji dengan 5-*cross validation* untuk meminimalisasi terjadinya *overfittng* dan *underfittng* pada model *machine learning*. Model dengan error paling sedikit akan dipilih nilai *alphanya* untuk melakukan prediksi.

4. **Seandainya dalam proses prediksi penugasan *problemset* diperbolehkan menambahkan data eksternal, apakah Anda akan menggunakan data eksternal? Jika iya, data apa yang akan Anda gunakan dan jelaskan alasannya! (NB: selain data primer harga laptop dengan spesifikasi yang sama, contoh: data harga laptop di marketplace)**

   Jika diperbolehkan menggunakan data eksternal, penulis akan menggunakan dataset semua harga primer beserta spesifikasi laptop yang ada di pasaran untuk melakukan imputasi pada nilai yang hilang pada dataset. Hal ini dilakukan penulis untuk menjaga nilai variansi pada *training data* tidak terlalu jauh dari kondisi awalnya. Selain itu, imputasi pada *training set* meningkatkan jumlah informasi dalam dataset. Peningkatan jumlah informasi dalam data akan membantu model menemukan pola dan *insight* dalam data.

5. **Bagaimana tanggapan dan evaluasi Anda terhadap problem set pada penugasan praktek dan soal teori pada proses open recruitment ini?**

   Menurut Penulis, Panitia OPREC sudah memikirkan dengan matang mengenai problem set baik dalam tugas praktek maupun tugas soal teori. Hal ini dapat terlihat ketika panitia menyusun tugas sedemikian rupa sehingga bisa diselesaikan dalam jangka waktu kurang dari dua minggu sebagaimana waktu seleksi dilaksanakan. Sejujurnya, tugas ini membuat penulis mempelajari algoritma dan matematika regresi lebih dalam lagi. Penulis ditantang untuk mempelajari lebih dalam mengenai dasar matematika regresi dan regularisasi yang saya aplikasikan untuk membuat notebook

yang akan saya kumpulkan bersama file ini. Penulis juga merasakan betapa kurangnya pengetahuan penulis mengenai teori probabilitas, turunan matriks dan aljabar linear. Setelah seleksi ini, penulis menjadi berkeinginan untuk memperdalam pemahaman penulis dalam statistika, *data mining*, kalkulus, dan aljabar linear. Akhir kata, Penulis berterima kasih kepada panitia yang sudah memberikan pengalaman seleksi yang baik dalam seleksi ini. Terima kasih.