Jawaban Tugas Seleksi Asisten Lab AI ‘23

Dipersiapkan Oleh Brian A. Hadian (13523048)
Versi: 1.0 07/08/2025

Jawaban

Modelling

Jawablah pertanyaan-pertanyaan berikut:

1. Jelaskan apa yang dimaksud dengan *hold-out validation* dan *k-fold cross-validation*!
   Validation merupakan teknik untuk memastikan bahwa model AI yang sedang mempelajari suatu dataset tertentu tidak memiliki kecenderungan untuk memenuhi secara tepat (overfitting) dan/atau tidak memiliki kecenderungan untuk memenuhi secara tepat terhadap keseluruhan data pada dataset tersebut (underfitting). Hal ini diperlukan agar model AI yang dilatih dapat menyesuaikan dengan data baru yang mungkin tidak sesuai / tidak dapat direpresentasikan oleh dataset yang dilatih.
   Hold out validation merupakan teknik validasi dengan membagi dataset menjadi dua bagian : training dataset dan test dataset. Training dataset digunakan untuk mengembangkan nilai weight (w) dan bias (b) dari sebelumnya diinisialisasi oleh bilangan kecil acak. Sementara itu, test dataset digunakan untuk menguji apakah model tersebut memiliki kemampuan untuk menyesuaikan terhadap dataset yang tidak berasal dari sesi *training*. Pembagian dataset menjadi biasanya disesuaikan dengan jumlah training data lebih banyak dari test data (~80 : ~ 20) sehingga dapat memberikan kesempatan untuk model AI meningkatkan akurasi dari weight dan bias. Dengan menggunakan teknik ini, model AI diharapkan dapat meningkatkan akurasi sehingga dapat memperoleh nilai yang lebih mendekati terhadap kondisi yang direpresentasikan oleh test dataset tersebut.
   Sementara itu, k-fold cross-validation merupakan teknik validasi dengan membagi dataset menjadi k bagian (subdataset), kemudian untuk salah satu subdataset tersebut akan berperan sebagai test dataset, sementara lainnya menjadi training dataset. Setelah menggunakan training dataset untuk melatih model AI, model tersebut akan menerima test dataset yang sudah dipilih sebelumnya untuk meningkatkan akurasi dari weight dan bias terkini. Kemudian, test dataset selanjutnya akan ditambahkan pada training dataset, dan diambil dataset

berbeda pada training dataset yang belum pernah digunakan sebagai test dataset. Selanjutnya, dilakukan kembali training menggunakan training dataset dan peningkatan akurasi menggunakan test dataset. Hal ini dilakukan terus menerus hingga seluruh subdataset telah berperan sebagai test dataset. Harapan dari teknik ini adalah model AI dapat menerima lebih banyak variasi sehingga dapat menyesuaikan nilai weight (w) dan bias (b) dengan lebih baik dan dapat melakukan prediksi dengan lebih akurat.

2. Jelaskan kondisi yang membuat *hold-out validation* lebih baik dibandingkan dengan *k-fold cross-validation*, dan jelaskan pula kasus sebaliknya!

   Kasus yang membuat hold out validation lebih baik adalah ketika dataset yang digunakan berjumlah sangat besar sehingga proses k-fold cross-validation dapat memakan komputasi yang relatif besar. Selain itu, hold-out validation lebih baik ketika proses komputasi memiliki biaya yang relatif mahal (ex. perkalian matriks dengan ukuran besar, dsb) sehingga dibutuhkan validasi yang lebih hemat.

   Sementara itu, kasus yang membuat k-fold cross-validation lebih baik adalah ketika dataset berjumlah sedikit dan variansi yang dimiliki oleh dataset tersebut cenderung kecil. Hal ini dapat diatasi oleh k-fold cross-validation karena k-fold dapat melatih model AI tersebut dengan memberikan dataset secara acak dan test dataset yang berganti-ganti. Selain itu, k-fold cross-validation lebih sensitif terhadap nilai weight (w) dan bias (b) sehingga dapat digunakan untuk melakukan tuning.

3. Apa yang dimaksud dengan *data leakage*?

   Data leakage adalah peristiwa suatu data terdapat pada suatu dataset lain yang digunakan sebagai test dataset. peristiwa ini dapat terjadi apabila tidak dilakukan preprocessing data untuk menghilangkan duplikasi

4. Bagaimana dampak *data leakage* terhadap kinerja dari model?

   Dampak dari data leakage adalah kinerja dari model yang cenderung tidak bisa memberikan prediksi yang baik apabila diberikan suatu masukan (input) yang tidak direpresentasikan pada test dataset maupun training dataset. Hal ini juga mengakibatkan model hanya bisa memberikan nilai yang baik untuk rentang yang diberikan oleh training dan test dataset yang digunakan (overfitting). Selain itu, hal ini juga mempersulit pembaharuan nilai weight (w) dan bias (b) apabila data leakage terjadi dalam ukuran besar

5. Berikanlah solusi untuk mengatasi permasalahan *data leakage*!
   a. melakukan penghapusan fitur (atribut) yang mengakibatkan data leakage

b. melakukan pelatihan ulang terhadap model dengan menggunakan dataset baru / dataset yang sudah dibersihkan
c. melakukan pembagian data ulang sehingga variansi yang dimiliki bisa bernilai cukup besar dibandingkan nilai variansi pada dataset sebelumnya