Bintang Nabiil Lukman
1103223126
TK-46-06

Analisa – RegresiUTSTelkom

1. **Jika model linear regression atau decision tree mengalami underfitting pada dataset ini, strategi apa yang akan digunakan untuk meningkatkan performanya? Bandingkan setidaknya dua pendekatan berbeda (misal: transformasi fitur, penambahan features, atau perubahan model ke algoritma yang lebih kompleks), dan jelaskan bagaimana setiap solusi memengaruhi bias-variance tradeoff!**

   Jawaban:
   Jika model linear regression atau decision tree mengalami underfitting (yaitu model tidak mampu menangkap pola dalam data), terdapat beberapa strategi yang bisa digunakan:
   a. Transformasi Fitur
   Melakukan transformasi non-linear pada fitur (misalnya: menambahkan fitur polinomial, logaritma, akar kuadrat, dsb) dapat membantu model sederhana seperti linear regression menangkap hubungan non-linear. Sebagai contoh, menambahkan fitur kuadrat dari variabel input bisa membuat model linear mampu belajar pola melengkung.
   - Dampak terhadap bias-variance tradeoff: Mengurangi bias karena model menjadi lebih fleksibel, tapi bisa sedikit meningkatkan variance jika terlalu banyak fitur non-linear ditambahkan.

   b. Mengganti ke Model yang Lebih Kompleks
   Jika decision tree mengalami underfitting karena terlalu dangkal (misalnya max_depth terlalu kecil), kita bisa menambah kedalamannya atau bahkan beralih ke model ensemble seperti Random Forest atau Gradient Boosting yang lebih kompleks.
   - Dampak terhadap bias-variance tradeoff: Mengurangi bias karena model bisa mempelajari lebih banyak detail dalam data. Namun, ada risiko overfitting (tingginya variance), terutama jika tidak dilakukan regularisasi.

2. **Selain MSE, jelaskan dua alternatif loss function untuk masalah regresi (misal: MAE, Huber loss) dan bandingkan keunggulan serta kelemahannya. Dalam skenario apa setiap loss function lebih cocok digunakan? (Contoh: data dengan outlier, distribusi target non-Gaussian, atau kebutuhan interpretasi model).**

Jawaban:

a. Mean Absolute Error (MAE)

Keunggulan:

- Lebih robust terhadap outlier karena tidak mengkuadratkan error.
- Lebih mudah diinterpretasikan karena satuannya sama dengan target.

Kelemahan:

- Tidak dapat diturunkan secara analitik di titik minimum (tidak diferensiabel di nol), sehingga bisa memperlambat konvergensi.

Cocok untuk: Data dengan banyak outlier atau jika kita ingin menghindari penalti besar dari error kuadrat.

b. Huber Loss

Kombinasi dari MSE dan MAE: menggunakan MSE untuk error kecil, dan MAE untuk error besar (terkendali oleh hyperparameter delta).

Keunggulan:

- Stabil dan robust terhadap outlier seperti MAE, namun tetap memiliki sifat diferensiabel seperti MSE.

Kelemahan:

- Perlu memilih nilai delta yang tepat.

Cocok untuk: Situasi di mana terdapat beberapa outlier, namun mayoritas data tidak terlalu bising.

3. **Tanpa mengetahui nama fitur, metode apa yang dapat digunakan untuk mengukur pentingnya setiap fitur dalam model? Jelaskan prinsip teknikal di balik metode tersebut (misal: koefisien regresi, feature importance berdasarkan impurity reduction) serta keterbatasannya!**

Jawaban:

a. Koefisien Regresi (Linear Model)

- Dalam linear regression, besar kecilnya nilai absolut dari koefisien menunjukkan pengaruh relatif suatu fitur.
- Prinsip teknikal: Koefisien merepresentasikan perubahan output terhadap perubahan unit fitur (dengan asumsi semua fitur distandarisasi).
- Keterbatasan:
  - Hanya valid untuk model linier dan asumsi multikolinearitas rendah.

b. Feature Importance dari Decision Tree

- Decision tree menghitung importance berdasarkan total pengurangan impurity (misalnya MSE) pada tiap node yang membelah fitur tersebut.
- Prinsip teknikal: Semakin besar pengurangan MSE akibat pembelahan suatu fitur, semakin penting fitur tersebut.

- Keterbatasan:
    - Bias terhadap fitur dengan banyak kategori atau skala numerik besar.
    - Tidak selalu stabil—hasil bisa berubah dengan data yang sedikit berbeda.

4. **Bagaimana mendesain eksperimen untuk memilih hyperparameter optimal (misal: learning rate untuk SGDRegressor, max_depth untuk Decision Tree) pada dataset ini? Sertakan analisis tradeoff antara komputasi, stabilitas pelatihan, dan generalisasi model!**

   Jawaban:

   a. Gunakan Grid Search atau Random Search
   - Grid Search: Menguji semua kombinasi dari set nilai hyperparameter.
   - Random Search: Memilih kombinasi secara acak—lebih efisien jika ruang hyperparameter besar.

   b. Evaluasi dengan K-Fold Cross Validation

   Membagi data ke dalam k bagian, melatih pada k-1 dan menguji pada sisanya, lalu merata-ratakan hasil.

   Tradeoff:
   - Komputasi: Grid search dengan cross-validation sangat mahal secara waktu, terutama jika dataset besar.
   - Stabilitas pelatihan: Cross-validation memberikan estimasi performa yang stabil dan menghindari overfitting pada satu subset.
   - Generalisasi: Dengan validasi silang, model lebih mampu memprediksi data baru karena diuji pada data yang tidak terlihat.

   c. Gunakan Early Stopping (jika model mendukung)

   Untuk model seperti Gradient Boosting, gunakan early stopping untuk menghindari overfitting saat training.

5. **Jika menggunakan model linear regression dan residual plot menunjukkan pola non-linear serta heteroskedastisitas, langkah-langkah apa yang akan diambil? (contohnya: Transformasi data/ubah model yang akan dipakai/etc)**

   Jawaban:

   a. Transformasi Fitur atau Target

   Gunakan transformasi Box-Cox, log, atau sqrt pada fitur atau target untuk mengurangi non-linearitas dan heteroskedastisitas.

   b. Gunakan Model Non-Linear

Alih-alih linear regression, gunakan decision tree, random forest, atau model berbasis neural network untuk menangkap pola non-linear.

c. Gunakan Generalized Linear Models (GLM)

GLM dengan fungsi link yang sesuai dapat menangani non-linearity dan varians yang tidak konstan.

d. Weighted Least Squares

Jika heteroskedastisitas tetap ada, weighted least squares memberi bobot berbeda pada observasi sehingga residu lebih seimbang.