Bias dan Varians dalam Machine Learning

Bias dan varians adalah dua sumber kesalahan utama dalam model machine learning. Memahami kedua konsep ini akan membantu Anda membangun model yang lebih baik dan menghindari masalah overfitting dan underfitting.

# Bias

Bias adalah kesalahan yang disebabkan oleh asumsi yang salah atau terlalu sederhana dalam model. Model yang bias memiliki kinerja yang buruk baik pada data latih maupun data uji.

## Contoh Bias:

* Menggunakan model regresi linear untuk data yang memiliki hubungan non-linear.
* Mengabaikan fitur penting yang sebenarnya mempengaruhi variabel target.

## Ciri-ciri Model dengan Bias Tinggi:

* Kesalahan prediksi tinggi pada data latih dan data uji.
* Model tidak dapat menangkap pola kompleks dalam data.

# Varians

Varians adalah kesalahan yang disebabkan oleh sensitivitas model terhadap fluktuasi kecil dalam data latih. Model yang memiliki varians tinggi akan memiliki kinerja yang baik pada data latih, tetapi buruk pada data uji.

## Contoh Varians

* Menggunakan model pohon keputusan yang terlalu dalam, sehingga model menghafal detail kecil dari data latih.
* Menggunakan model dengan terlalu banyak parameter yang tidak diperlukan.

## Ciri-ciri Model dengan Varians Tinggi

* Kesalahan prediksi rendah pada data latih, tetapi tinggi pada data uji.
* Model terlalu kompleks dan cenderung overfitting pada data latih.

# Trade-off antara Bias dan Varians

Terdapat trade-off antara bias dan varians dalam machine learning. Model yang kompleks cenderung memiliki bias rendah tetapi varians tinggi, sedangkan model yang sederhana cenderung memiliki bias tinggi tetapi varians rendah.

## Tujuan:

Menemukan keseimbangan yang optimal antara bias dan varians untuk mendapatkan model yang dapat menggeneralisasi dengan baik pada data baru.

## Cara Mengatasi Bias Tinggi:

* Menggunakan model yang lebih kompleks.
* Menambahkan lebih banyak fitur.
* Mengurangi regularisasi.

## Cara Mengatasi Varians Tinggi:

* Menggunakan model yang lebih sederhana.
* Mengurangi jumlah fitur.
* Menambahkan regularisasi.

**Tugas Hari 31:**

1. **Pemahaman Konsep:** Jelaskan dengan kata-kata Anda sendiri apa itu bias dan varians dalam machine learning. Berikan contoh masing-masing.
2. **Analisis Model:** Amati model-model yang telah Anda buat sebelumnya (regresi linear, regresi polinomial, SVM). Menurut Anda, model mana yang cenderung memiliki bias tinggi dan model mana yang cenderung memiliki varians tinggi? Jelaskan alasan Anda.
3. **Eksperimen (Opsional):** Coba ubah kompleksitas model (misalnya, dengan mengubah derajat polinomial dalam regresi polinomial atau parameter C dalam SVM). Amati bagaimana perubahan ini mempengaruhi bias dan varians model.

**Selamat Mengerjakan Tugas! 🙂**

**Tugas**

1. **Pemahaman Konsep**

**Bias:** Bias itu terjadi karena adanya prediksi yang lebih tinggi dari perkiraan kita.

*Contoh:* Saya memiliki sebuah data tahunan dan saya akan melakukan prediksi terhadap data tersebut. Pada saat itulah *Bias* terjadi ketika melakukan prediksi yang diluar perkiraan saya.

**Varians:** Varians itu bisa terjadi karena adanya saya melakukan data latih secara terus-menerus, sehingga data tersebut sudah menghafalkan data saya dan terjadinya overfitting pada data saya.

*Contoh:* Saya memiliki sebuah sample data yang dimana sample tersebut merupakan data latih, data latih ini saya mencoba melakukan berbagai latihan dengan adanya beberapa model sehingga terjadinya overfitting pada data sample saya atau data latih saya.

1. **Analisis Model**

Regresi linear mungkin memiliki Bias yang tinggi. Alasannya karena Model tersebut cenderung melakukan Penghafal pada data latih.

1. **Eksperimen**

# Umpan Balik:

* **Pemahaman Konsep:** Penjelasan Anda tentang bias dan varians sudah tepat. Bias adalah kesalahan yang disebabkan oleh asumsi model yang salah atau terlalu sederhana, sedangkan varians adalah kesalahan yang disebabkan oleh sensitivitas model terhadap fluktuasi kecil dalam data latih. Contoh yang Anda berikan juga menggambarkan dengan baik bagaimana bias dan varians dapat terjadi dalam praktiknya.
* **Analisis Model:** Anda menyatakan bahwa regresi linear cenderung memiliki bias tinggi. Namun, perlu diperhatikan bahwa regresi linear itu sendiri tidak selalu memiliki bias tinggi. Bias tinggi akan terjadi jika model regresi linear digunakan untuk memodelkan data yang memiliki hubungan non-linear. Jika hubungan antara variabel independen dan dependen benar-benar linear, maka regresi linear seharusnya tidak memiliki bias tinggi.

**Klarifikasi:**

* **Bias Tinggi pada Regresi Linear:** Regresi linear akan memiliki bias tinggi jika digunakan untuk memodelkan data yang sebenarnya memiliki hubungan non-linear antara variabel independen dan dependen. Misalnya, jika kita mencoba memodelkan hubungan antara usia dan pendapatan dengan regresi linear, padahal sebenarnya hubungan tersebut berbentuk kurva, maka model akan memiliki bias tinggi karena tidak dapat menangkap pola non-linear tersebut.
* **Overfitting pada Regresi Linear:** Overfitting pada regresi linear lebih mungkin terjadi jika model memiliki terlalu banyak fitur atau terlalu kompleks dibandingkan dengan jumlah data yang tersedia. Hal ini dapat menyebabkan model menghafal noise atau fluktuasi acak dalam data latih, bukannya belajar pola umum yang mendasarinya.

**Saran Perbaikan:**

* **Analisis Model:** Untuk menentukan apakah model regresi linear memiliki bias tinggi, Anda perlu memeriksa apakah ada pola non-linear dalam data Anda. Anda dapat melakukan ini dengan membuat scatter plot antara variabel independen dan dependen. Jika Anda melihat pola non-linear, maka regresi linear mungkin bukan model yang tepat.
* **Eksperimen:** Untuk melihat bagaimana perubahan kompleksitas model mempengaruhi bias dan varians, Anda dapat mencoba beberapa hal berikut:
  + **Regresi Polinomial:** Jika Anda melihat pola non-linear dalam data, coba gunakan regresi polinomial dengan derajat yang lebih tinggi.
  + **Regularisasi:** Coba tambahkan regularisasi (L1 atau L2) pada model regresi linear Anda. Ini dapat membantu mengurangi overfitting dan varians.
  + **Cross-Validation:** Gunakan cross-validation untuk mengevaluasi kinerja model dengan berbagai tingkat kompleksitas dan hyperparameter. Ini akan membantu Anda menemukan model yang memiliki keseimbangan terbaik antara bias dan varians.

**Kesimpulan:**

Memahami konsep bias dan varians sangat penting dalam membangun model machine learning yang baik. Dengan melakukan analisis dan eksperimen yang tepat, Anda dapat mengidentifikasi dan mengatasi masalah bias dan varians dalam model Anda.