Tugas perbaikan

Bab 2

Nama : muhammad makhlufi makbullah

Kelas : TK 45 01

NIM : 1103210171

1. Apa itu Supervised Learning?

Supervised learning adalah metode pembelajaran mesin di mana model dilatih menggunakan data yang telah dilabeli. Dalam hal ini, data latih terdiri dari pasangan input-output yang diketahui. Model berusaha mempelajari hubungan antara input dan output sehingga dapat memprediksi output untuk data baru yang belum pernah dilihat sebelumnya.

- **Contoh Klasifikasi:** Menentukan apakah sebuah email adalah spam atau bukan spam berdasarkan fitur-fitur seperti kata kunci, panjang pesan, dan pengirim.
- **Contoh Regresi:** Memprediksi harga rumah berdasarkan fitur seperti ukuran rumah, jumlah kamar, dan lokasi.

---

2. Klasifikasi vs Regresi

Supervised learning dapat dibagi menjadi dua kategori utama:

- **Klasifikasi:** Model memprediksi label kategori untuk data. Misalnya, memprediksi apakah email adalah spam atau bukan spam. Labelnya bisa berupa kategori seperti "setosa", "versicolor", atau "virginica" pada dataset Iris.

Contoh algoritma klasifikasi:

    - K-Nearest Neighbors (k-NN)
    - Decision Trees
    - Support Vector Machines (SVM)
    - Logistic Regression
- **Regresi:** Model memprediksi nilai kontinu. Misalnya, memprediksi harga rumah berdasarkan fitur seperti ukuran rumah, jumlah kamar tidur, dan lokasi.

Contoh algoritma regresi:

    - Linear Regression
    - Ridge Regression
    - Lasso Regression

---

3. Fungsi Loss

Dalam supervised learning, model berusaha meminimalkan *loss function* atau fungsi kerugian. Fungsi loss mengukur perbedaan antara prediksi model dan nilai yang sebenarnya. Tujuannya adalah untuk membuat model yang meminimalkan nilai loss ini, sehingga prediksi model semakin mendekati kenyataan.

- **Contoh fungsi loss:**
    - **Klasifikasi:** *Cross-entropy loss* atau *log loss*.
    - **Regresi:** *Mean squared error* (MSE).

---

4. Algoritma Supervised Learning

Bab ini mengenalkan beberapa algoritma utama yang digunakan dalam supervised learning untuk klasifikasi dan regresi. Beberapa algoritma tersebut meliputi:

- **K-Nearest Neighbors (k-NN):** Algoritma ini mengklasifikasikan data berdasarkan kedekatannya dengan data lain. Misalnya, untuk mengklasifikasikan bunga Iris, model akan memeriksa tetangga terdekat dari sebuah titik data dan memberi label berdasarkan mayoritas tetangga tersebut.
- **Logistic Regression:** Digunakan untuk klasifikasi biner (misalnya, spam vs. non-spam). Walaupun namanya mengandung "regression", ini sebenarnya adalah algoritma klasifikasi.
- **Support Vector Machines (SVM):** Algoritma yang menemukan batas optimal antara kelas-kelas yang berbeda dengan memaksimalkan margin antara data.
- **Decision Trees:** Model pohon yang membuat keputusan berdasarkan serangkaian aturan. Setiap cabang mewakili keputusan berdasarkan fitur data, dan setiap daun berisi label prediksi.
- **Random Forests:** Kombinasi dari beberapa decision trees yang digunakan untuk meningkatkan akurasi dan mengurangi overfitting.
- **Linear Regression:** Digunakan dalam regresi untuk memprediksi nilai kontinu berdasarkan hubungan linear antara variabel input dan output.

---

5. Overfitting dan Underfitting

- **Overfitting:** Ketika model terlalu rumit dan terlalu mempelajari detail dari data latih, sehingga tidak generalisasi dengan baik pada data uji. Model "terlalu fit" dengan data latih dan gagal menangani data baru.
- **Underfitting:** Ketika model terlalu sederhana dan tidak cukup belajar dari data latih, sehingga tidak dapat menangkap pola yang ada. Model gagal untuk memodelkan data dengan baik.

Untuk menghindari masalah ini, kita perlu menggunakan teknik seperti *cross-validation*, pengaturan parameter model, dan pengaturan kompleksitas model (misalnya, mengurangi kedalaman decision tree).

---

6. Evaluasi Model

Setelah model dilatih, kita perlu mengevaluasi kinerjanya menggunakan data uji. Beberapa metrik evaluasi yang digunakan dalam supervised learning adalah:

- **Akurasi:** Persentase prediksi yang benar untuk klasifikasi.
- **Precision dan Recall:** Digunakan untuk mengukur kinerja model klasifikasi, terutama dalam kasus ketidakseimbangan kelas (misalnya, ketika satu kelas jauh lebih sering muncul dari kelas lainnya).
- **F1-score:** Rata-rata harmonis antara precision dan recall.
- **Mean Squared Error (MSE):** Digunakan dalam regresi untuk mengukur seberapa besar prediksi model berbeda dengan nilai yang sebenarnya.

---

7. Mengatasi Tantangan dalam Supervised Learning

- **Pembersihan Data:** Data mungkin mengandung nilai yang hilang atau tidak konsisten. Membersihkan dan menyiapkan data dengan baik sangat penting untuk kinerja model.
- **Pemilihan Fitur:** Memilih fitur yang relevan dan mengurangi fitur yang tidak penting dapat meningkatkan kinerja model dan mengurangi risiko overfitting.
- **Skalabilitas:** Beberapa algoritma, seperti SVM, dapat memakan waktu jika dataset sangat besar, sehingga diperlukan teknik untuk meningkatkan skalabilitas.

poin penting yang ada dalam Bab 2: Supervised Learning

1. Classification and Regression

- **Classification**: Tugas di mana model memprediksi kategori atau label diskrit (misalnya, mengklasifikasikan email sebagai spam atau tidak spam).
- **Regression**: Tugas di mana model memprediksi nilai kontinu (misalnya, memprediksi harga rumah berdasarkan beberapa fitur).

---

2. Generalization, Overfitting, and Underfitting

- **Generalization**: Kemampuan model untuk bekerja dengan baik pada data baru yang belum dilihat sebelumnya.
- **Overfitting**: Ketika model terlalu kompleks dan mempelajari noise atau detail yang tidak relevan dalam data latih, sehingga kinerjanya buruk pada data uji.
- **Underfitting**: Ketika model terlalu sederhana untuk menangkap pola dalam data, sehingga tidak mampu menghasilkan prediksi yang akurat.

---

3. Relation of Model Complexity to Dataset Size

- **Model Complexity**: Semakin kompleks model (misalnya, semakin dalam decision tree), semakin besar risiko overfitting, terutama pada dataset kecil.
- **Dataset Size**: Dengan dataset yang lebih besar, model yang lebih kompleks bisa lebih efektif karena lebih banyak data membantu model belajar pola yang lebih umum.

---

4. Supervised Machine Learning Algorithms

Bab ini membahas berbagai algoritma pembelajaran mesin yang digunakan dalam supervised learning untuk tugas klasifikasi dan regresi, seperti:

- **k-Nearest Neighbors (k-NN)**
- **Linear Models (misalnya, Linear Regression)**
- **Naive Bayes**

- **Decision Trees**
- **Ensembles of Decision Trees (misalnya, Random Forests)**
- **Support Vector Machines (SVM)**

---

5. Some Sample Datasets

Beberapa dataset yang sering digunakan dalam supervised learning termasuk:

- **Iris dataset**: Digunakan untuk klasifikasi bunga Iris berdasarkan fitur morfologi.
- **Boston Housing dataset**: Digunakan untuk regresi, memprediksi harga rumah berdasarkan fitur tertentu.
- **Digits dataset**: Digunakan untuk klasifikasi gambar digit tulisan tangan.

---

6. k-Nearest Neighbors (k-NN)

- Algoritma klasifikasi yang sederhana namun efektif. Model mengklasifikasikan data berdasarkan kedekatannya dengan titik data lainnya (tetangga terdekat).
- **Kelebihan**: Mudah dipahami dan diimplementasikan, serta tidak memerlukan pelatihan eksplisit.
- **Kekurangan**: Bisa sangat lambat pada dataset besar karena memerlukan perhitungan jarak antar titik untuk setiap prediksi.

---

7. Linear Models

- **Linear Regression**: Digunakan untuk regresi, memodelkan hubungan linear antara fitur dan target.
- **Logistic Regression**: Digunakan untuk klasifikasi biner, meskipun namanya mengandung "regression", ini adalah model klasifikasi.

- **Kelebihan**: Mudah dipahami dan diinterpretasikan, cepat dihitung.
- **Kekurangan**: Tidak bisa menangani hubungan non-linear yang kompleks tanpa pemrosesan lebih lanjut.

---

8. Naive Bayes Classifiers

- **Naive Bayes**: Algoritma klasifikasi berbasis probabilitas yang mengasumsikan independensi antar fitur.
- **Kelebihan**: Cepat dan efisien untuk dataset besar, baik untuk klasifikasi teks (misalnya, spam filtering).
- **Kekurangan**: Asumsi independensi sering kali tidak benar, yang dapat mempengaruhi kinerja pada data nyata.

---

9. Decision Trees

- **Decision Tree**: Algoritma yang membuat keputusan berdasarkan serangkaian aturan, yang membentuk struktur pohon.
- **Kelebihan**: Mudah diinterpretasikan dan tidak memerlukan praproses data.
- **Kekurangan**: Rentan terhadap overfitting, terutama jika pohon terlalu dalam.

---

10. Ensembles of Decision Trees

- **Random Forests**: Penggunaan beberapa decision tree untuk meningkatkan akurasi dan mengurangi overfitting dengan mengambil voting dari semua pohon.
- **Boosting (misalnya, AdaBoost, Gradient Boosting)**: Algoritma ensemble yang membangun model secara bertahap, memperbaiki kesalahan dari model sebelumnya.

---

11. Kernelized Support Vector Machines (SVM)

- **SVM dengan Kernel**: SVM mengklasifikasikan data dengan menemukan hyperplane pemisah yang optimal. Dengan kernel, SVM dapat menangani data non-linear dengan memetakan data ke dimensi yang lebih tinggi.
- **Kelebihan**: Efektif untuk data dengan dimensi tinggi dan untuk masalah klasifikasi non-linear.
- **Kekurangan**: Memerlukan banyak waktu komputasi untuk dataset besar dan pemilihan kernel yang tepat.

---

12. Neural Networks (Deep Learning)

- **Neural Networks**: Terinspirasi oleh cara kerja otak manusia, neural networks memiliki lapisan-lapisan yang dapat mempelajari representasi data yang kompleks.
- **Deep Learning**: Menggunakan jaringan saraf dengan banyak lapisan tersembunyi untuk menangani tugas-tugas yang sangat kompleks seperti pengenalan gambar dan pemrosesan bahasa alami.
- **Kelebihan**: Sangat kuat untuk data yang sangat besar dan kompleks.
- **Kekurangan**: Membutuhkan banyak data dan sumber daya komputasi.

---

13. Uncertainty Estimates from Classifiers

- **Uncertainty Estimation**: Beberapa model dapat memberikan estimasi ketidakpastian tentang prediksi mereka. Ini membantu untuk mengetahui seberapa percaya diri model terhadap hasil prediksi.
- **Contoh**: Pada SVM, prediksi dapat memiliki margin ketidakpastian yang menunjukkan seberapa jauh data dari garis pemisah.

---

14. The Decision Function

- **Decision Function**: Fungsi yang digunakan oleh model klasifikasi untuk menentukan di kelas mana sebuah titik data termasuk berdasarkan fitur yang dimilikinya.

---

15. Predicting Probabilities

- Beberapa algoritma (misalnya, Logistic Regression, Naive Bayes) dapat memberikan prediksi probabilitas untuk setiap kelas, bukan hanya label yang paling mungkin. Ini berguna ketika Anda ingin mengetahui seberapa yakin model terhadap keputusan yang diambil.

---

16. Uncertainty in Multiclass Classification

- **Multiclass Classification**: Pada klasifikasi dengan lebih dari dua kelas, model perlu memprediksi lebih dari satu kelas dan menangani ketidakpastian dengan memberikan probabilitas untuk setiap kelas.

---

17. Summary and Outlook

- Bab ini menyimpulkan berbagai algoritma yang digunakan dalam supervised learning, menggarisbawahi pentingnya pemilihan algoritma yang tepat, serta tantangan seperti overfitting, underfitting, dan pemilihan fitur.
- **Outlook**: Meskipun banyak algoritma yang sudah terbukti efektif, riset dan perkembangan dalam pembelajaran mesin terus berkembang, khususnya dalam hal teknik deep learning dan pemrosesan data besar.