# Support Vector Machine (SVM)

# Support Vector Machine (SVM)

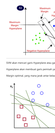
Website to output the latest interpretable guide online

## Apa itu SVM?

Support vector machine (SVM) adalah metode pembelajaran supervised yang digunakan proses set dari input yang telah terlatih (TRAIN) dan ditandai sebagai salah satu dari dua kategori dapat dibangun model SVM untuk memprediksi kategori input baru. Model SVM yang dihasilkan dipetakan sebuah contoh pelatihan menjadi titik ruang sehingga contoh dari kategori yang berbeda dibagi dengan celah jelas yang lebar. Contoh-contoh baru kemudian dipetakan ke ruang yang sama dan diprediksi termasuk ke kategori mana berdasarkan sisi celah mereka berada.

- SVM untuk Klasifikasi dasar Support Vector Classifier
- SVM untuk Regresi dasar Support Vector Regression

## Support Vector Machine




SVM dapat melakukan garis hyperplane dua garis pemisah untuk klasifikasi
Hyperplane dan sebuah garis pemisah yang bisa melakukan regresi

Mengap pemisah yang bisa kita pilih dan mengapa kita ingin kita coba tebak?

## Contoh

Tentukan hyperplane Kasus/kasus pengambilan keputusan yang bisa / kemungkinan data untuk kelas.

## Other possible solutions

## Pertanyaan

1. Kelas yang lebih baik adalah B atau A?

2. Bagaimana cara anda menentukan yang lebih baik?

## Jawaban

Ini mana saya ini ?

## Hard Margin vs Soft Margin

Margin adalah jarak dari suatu nilai dengan hyperplane

- Hard Margin: Sangat ketat yang mengus ung mampu memisahkan kelas secara sempurna
- Soft Margin: Sangat lunak yang mau ini memberikan ruang untuk kesalahan, sehingga dapat menyelesaikan overfitting

Hard Margin
Soft Margin
Margin
Margin
Support vectors
Support vectors
Support vectors

## Then, we have a tree like this now

**For this case, on the same data but all features and not just the top 3 as we did earlier**

## Until we have a complete the tree and do the classification

Bertambah lah1 yang yang lagi yang lainnya di atasnya lagi.

- Pic Cool=Yes, Pic Cool=Rainy, Pic Cool=Sunny, Pic Hot=Yes, Pic Hot=No
- Pic Cool=Yes, Pic Cool=Rainy, Pic Hot=Yes, Pic Hot=Yes

## Gini

$$Gini(x) = 1 - \sum_{j=1}^{n} (p_j)^2$$

| | | |
|---|---|---|
| C1 | 0 | $P(C1)=0/6$ $P(C2)=6/6$ |
| C2 | 6 | $Gini = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$ |
| | | |
| C1 | 3 | $P(C1)=3/6$ $P(C2)=3/6$ |
| C2 | 3 | $Gini = 1 - (1/2)^2 - (1/2)^2 = 1 - 1/2 - 1/2 = 0.5$ |
| | | |
| C1 | 2 | $P(C1)=2/6$ $P(C2)=4/6$ |
| C2 | 4 | $Gini = 1 - (1/3)^2 - (2/3)^2 = 1 - 1/9 - 4/9 = 0.44$ |

## Entropy

$$Entropy(x) = \sum_{j=1}^{n} -p_j \log_2(p_j)$$

| | | |
|---|---|---|
| C1 | 0 | $P(C1)=0/6=0$ $P(C2)=6/6=1$ |
| C2 | 6 | $Entropy = -(0\log_2 0 - 1\log_2 1) = -(0 - 0) = 0$ |
| | | |
| C1 | 3 | $P(C1)=3/6$ $P(C2)=3/6$ |
| C2 | 3 | $Entropy = -(0.5\log_2 0.5 + 0.5\log_2 0.5) = 1$ |
| | | |
| C1 | 2 | $P(C1)=2/6$ $P(C2)=4/6$ |
| C2 | 4 | $Entropy = -(0.3\log_2 0.3 + 0.6\log_2 0.6) = 0.92$ |

## Tree Pruning

Pemangkasan pada pohon keputusan merupakan proses untuk dalam data pelatihan akan erasa atau culler. Pohon yang dipangkas lebih kecil dari pada utuh.

Ada dua pendekatan untuk memangkas pohon:

1. Pra-pemangkasan

Pohon dipangkas dengan mengizinkan kesalahan pelatihan lebih besar

2. Pascapemangkasan Pendekatan ini dilakukan pada pohon lengkap yang sudah tumbuh

## Pro

1. Model dimengerti
2. Dapat menangani variabel filter
3. Cepat untuk pemrosesan data
4. Dapat menangani data besar
5. Efisiensi Nya Mumpuni

## Kontra

1. Cenderung Tidak Stabil, adalah salah kesalahan paling grafik untuk model pohon keputusan. Bisaseh ini disebabkan dengan pengurangan kecil saja yang dan ini bisa sama bahkan ini yang ini
2. Tidak cocok untuk variabel kontinu, Saat dipang dengan variabel numerik kontinu, pohon keputusan kehilangan informasi saat mengelompokkan variabel dalam kategori yang berbeda.

# Random Forest

## Apa Itu Random Forest?

Mengembangkan dari Decision Tree




## Bagging

Melakukan prediksi atau membangun suatu keputusan dengan cara membuat beberapa model Bootstrap yang sama




Setelah ada data baru, maka dilakukan voting untuk model 1 model 2 dan model 3. Hasil dibagaikan berdasarkan voting. Dengan cara voting yang terbanyak.

## Random Forest Work




Sebagai contoh, kumpulan Original dataset. Kemudian dibagi menjadi Creation Bootstrap Dataset yang yang sama yang bagian dipindah pengolog secara random sampling dan dan dimasukkan dengan pengembalian sampling yang disebut bagger 1 sampling untuk bagger 2 sampling untuk bagger 3. Kemudian masing masing bagger membuat decision tree.

## Random Forest Hyperparameter

- nip Bagian ini adalah jumlah traksimum variabel yang dibolehkan oleh Random Forest untuk dicoba di pohon tertentu. Maksudnya bisa jadi yang variabel ke dalam sampling. Misal ada 10 kolom. Maksimal kolomnya 5 maka yang dicobain pada pohon tertentu cuma 5 kolom aja.
- n_estimator. Ini adalah jumlah pohon yang ingin Anda bangun sebelum mengambil pemungutan suara maksimum atau rata-rata prediksi. Jumlah pohon yang lebih besar meningkatkan stabilitas dan kinerja prediksi.

# Pro

1. Keseimbangan Kesalah Jumlah pohon keputusan dan juga mengurangi ketem dari ketem ini. Meningkatkan akurasi
2. Random Forest dapat secara otomatis menangani data yang hilang
3. Random Forest dapat di gunakan untuk data yang besar dan kompleksitas tidak masa Random Forest bisa menggunakan pendekatan Bootstrap untuk data yang besar
4. Random Forest Gampang kan untuk scaler dan dapat menangani skala otomatis.

# Kontra

1. Kompleksitas. Random Forest memiliki banyak pohon Cobe hasil ada pohon dalam bisa pohon keputusan dan menggabungkan hasilnya. Bisa sebuah di manfaat 100 pohon di possible dalam Gohon. Cenderung kelambatan algoritma ini dalam memprediksi data baru karena harus melewati banyak pohon untuk membuat prediksi. Ini bisa menjadi masalah untuk tugas real-time dan kecepatan
2. Interpretasi Sulit. Meskipun Random Forest memberikan akurasi yang baik, sulit untuk dipahami dengan pohon keputusan tunggal. Model yang lebih kompleks dengan banyak pohon cenderung lebih sulit dimengerti dan dan pohon keputusan yang lebih sederhana dapat lebih mudah dipahami oleh manusia dan lebih mudah dijelaskan kepada pengguna.