

# EVALUASI MODEL

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

*aliakbars@live.com*

April 1, 2020

# SELAYANG PANDANG

① ULASAN

② GENERALISASI

③ OPTIMASI MODEL DARI DATASET

④ METRIK EVALUASI

## BAHAN BACAAN

- 1 VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media. <https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html>
- 2 Le Calonnec, Y. (October 2017). CS229 Bias-Variance and Error Analysis. <http://cs229.stanford.edu/section/error-analysis.pdf>
- 3 Ng, A. (2019). CS229 Lecture notes: Regularization and model selection. <http://cs229.stanford.edu/notes2019fall/cs229-notes5.pdf>

# ULASAN

## MATERI SEBELUMNYA...

- Regresi linear
- *Sum of squared error* dari *log likelihood*
- Transformasi fitur dan regularisasi

Simak video ini:  
Lecture 9 - CS229 Machine Learning (Stanford)

# GENERALISASI

# GENERALISASI ERROR

- Tujuan kita adalah menghasilkan model yang dapat bekerja baik pada **semua data**
- Tidak mungkin mendapatkan semua data
- Solusi: Gunakan **data latih** dan **data uji**



# GENERALISASI ERROR

- *Training data:*  $\{x_i, y_i\}$
- *Future data:*  $\{x_i, ?\}$
- Target: Model bekerja baik pada **future data**

Mengapa?

# OVERFITTING

- Model terlalu kompleks, **terlalu fleksibel**
- Mengenali dan memasukkan *noise* dari dalam data latih ke dalam model
- Mengenali pola yang *tidak akan muncul lagi*

# OVERFITTING: DEFINISI

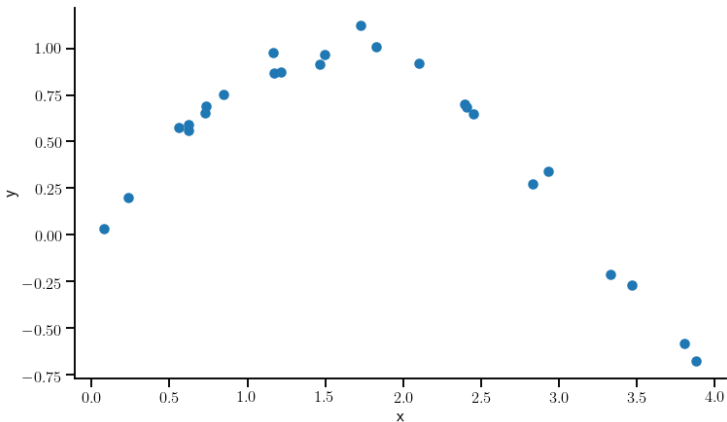
Model  $F$  dikatakan *overfitting* jika:

- 1 kita dapat menemukan model lain  $F'$
- 2 dengan error lebih besar pada data latih:  
$$E_{train}(F') > E_{train}(F)$$
- 3 tetapi error lebih kecil pada data uji:  $E_{gen}(F') < E_{gen}(F)$

# UNDERFITTING

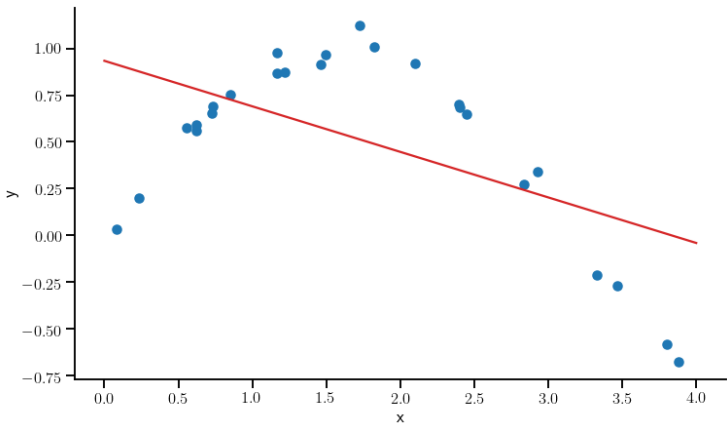
- Model terlalu kaku, **terlalu simpel**
- Tidak berhasil menemukan pola yang penting
- Masih ada model yang bisa menghasilkan  $E_{train}$  dan  $E_{gen}$  lebih rendah

## CONTOH PADA REGRESI



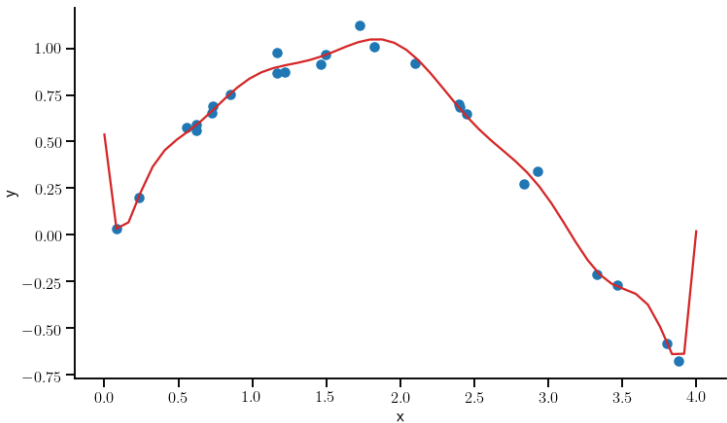
**GAMBAR:** Bagaimana kira-kira hasil regresi pada data seperti ini?

## CONTOH PADA REGRESI



GAMBAR: Regresi polinomial dengan  $p = 1$  (linear)

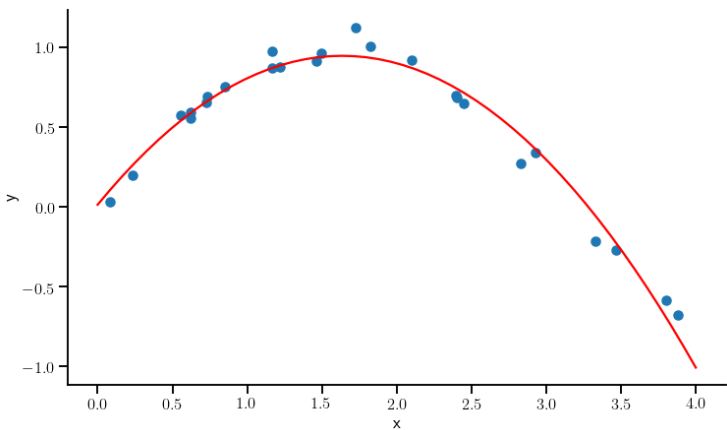
## CONTOH PADA REGRESI



GAMBAR: Regresi polinomial dengan  $p = 12$

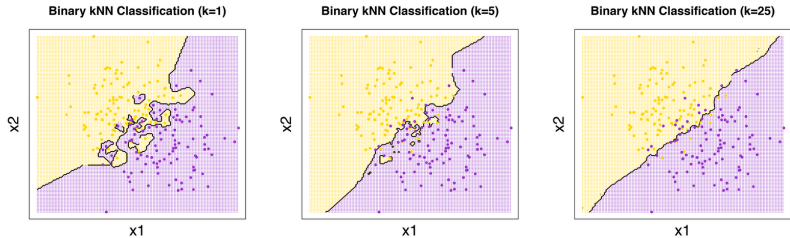


## CONTOH PADA REGRESI



GAMBAR: Regresi polinomial dengan  $p = 2$

# CONTOH PADA KLASIFIKASI



**GAMBAR:** Batas klasifikasi berubah seiring dengan perubahan nilai  $k$  (DeWilde, 2012)

# FLEKSIBILITAS PREDIKTOR

- Setiap dataset perlu prediktor dengan **fleksibilitas yang berbeda**, tergantung kesulitannya dan data yang tersedia

# FLEKSIBILITAS PREDIKTOR

- Setiap dataset perlu prediktor dengan **fleksibilitas yang berbeda**, tergantung kesulitannya dan data yang tersedia
- Diperlukan **kenop** untuk mengubah fleksibilitasnya, e.g.
  - regresi: orde polinomial
  - NB: jumlah atribut,  $\epsilon$
  - k-NN: nilai k

# FLEKSIBILITAS PREDIKTOR

- Setiap dataset perlu prediktor dengan **fleksibilitas yang berbeda**, tergantung kesulitannya dan data yang tersedia
- Diperlukan **kenop** untuk mengubah fleksibilitasnya, e.g.
  - regresi: orde polinomial
  - NB: jumlah atribut,  $\epsilon$
  - k-NN: nilai k
- Idenya, memutar kenop tersebut untuk **menghasilkan error yang rendah secara umum**

# ERROR LATIHAN VS GENERAL

- Error latihan:

$$E_{train} = \frac{1}{N} \sum_{i=1}^N error(f_D(\mathbf{x}_i), y_i)$$

- Error general:

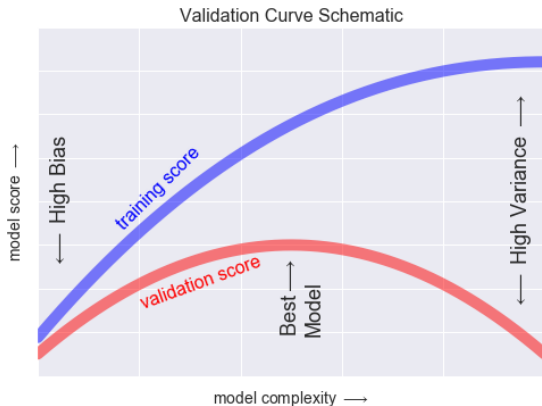
$$E_{gen} = \int error(f_D(\mathbf{x}), y) p(y, \mathbf{x}) d\mathbf{x}$$

- Kita hanya tahu **jangkauan** dari  $\{x, y\}$

Estimasi nilainya dengan

$$E_{test} = \frac{1}{N} \sum_{i=1}^N error(f_D(\mathbf{x}_i), y_i)$$

# BIAS-VARIANCE TRADE-OFF



**GAMBAR:** Perubahan nilai *metric* sesuai dengan kompleksitas model



## CONTOH KASUS

Dalam regresi linear:

- Apa yang harus diubah pada model untuk **mengurangi bias**?

# CONTOH KASUS

Dalam regresi linear:

- Apa yang harus diubah pada model untuk **mengurangi bias**?
- Bagaimana dengan **variansi**?

# CONTOH KASUS

Dalam regresi linear:

- Apa yang harus diubah pada model untuk **mengurangi bias**?
- Bagaimana dengan **variansi**?
- Pada **dataset** yang mana modelnya harus kita **evaluasi**?

# OPTIMASI MODEL DARI DATASET

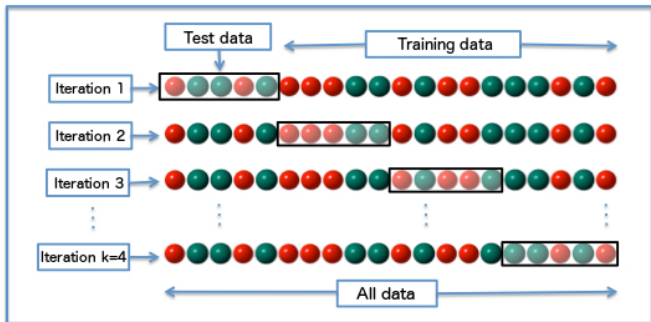
# TRAINING, VALIDATION, TESTING SETS

- **Data latih:** konstruksi *classifier*
- **Data validasi:** memilih algoritma dan *parameter tuning*
- **Data uji:** mengestimasi *error rate* secara umum
- Catatan: Bagi datanya secara **acak!**

# CROSS-VALIDATION

- Datanya kadang tidak cukup banyak untuk dibagi!
- Ide: latih dan uji secara bergantian
- Umumnya: 10-fold cross-validation

# CROSS-VALIDATION



GAMBAR: 4-fold cross-validation

# LEAVE-ONE-OUT

*n-fold cross-validation*

## PROS

Menghasilkan *classifier* terbaik

## CONS

- Ongkos komputasi tinggi: melatih model  $n$  kali
- Kelas tidak seimbang: untuk data terduplikasi, 1NN menghasilkan 0% error



# METRIK EVALUASI

# UNBALANCED DATASET

- e.g. Prediksi apakah akan terjadi gempa atau tidak!



# UNBALANCED DATASET

- e.g. Prediksi apakah akan terjadi gempa atau tidak!
- Jika selalu diklasifikan sebagai “tidak”, akurasi akan maksimal, error akan minimal.



# UNBALANCED DATASET

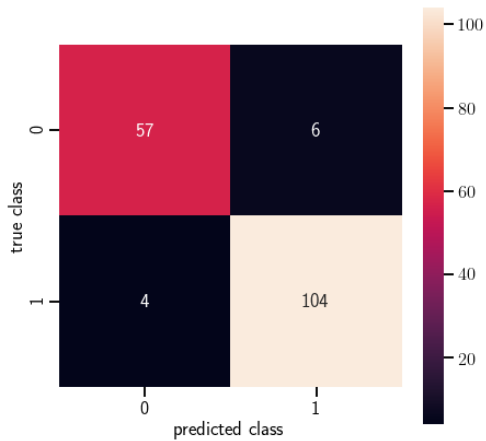
- e.g. Prediksi apakah akan terjadi gempa atau tidak!
- Jika selalu diklasifikan sebagai “tidak”, akurasi akan maksimal, error akan minimal.
- Solusi: Gunakan metrik lain



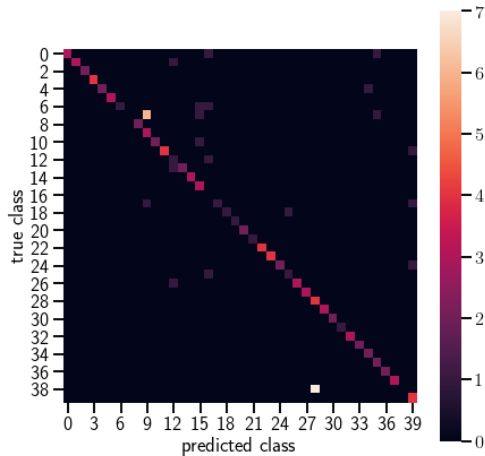
## MISSES & FALSE ALARMS

- False Alarm rate = False Positive rate =  $FP/(FP + TN)$
- Miss rate = False Negative rate =  $FN/(TP + FN)$
- Recall = True Positive rate = Sensitivity =  $TP/(TP + FN)$
- Precision =  $TP/(TP + FP)$
- Specificity =  $1 - FPR = TN/(TN + FP)$
- Harus dilaporkan berpasangan!

# CONFUSION MATRIX



# CONFUSION MATRIX - MULTICLASS



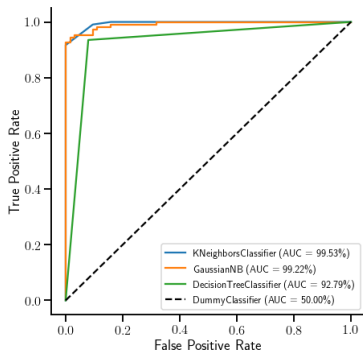
# UTILITY & COST

- Terkadang perlu satu angka untuk pembandingan antarmodel
- **Detection cost:**  $cost = c_{FP} \times FP + c_{FN} \times FN$
- **F-measure:**  $F_1 = 2 \times \frac{precision \times recall}{precision + recall}$



# ROC CURVES

Receiver Operating Characteristic: TPR vs FPR dengan perubahan *threshold*



Menghitung Area Under the Curve (AUC) sebagai pengganti akurasi

## CONTOH

Asumsikan Anda mempunyai 4 contoh dengan kelas positif (+1) dan 8 contoh dengan kelas negatif (-1). Anggaplah bahwa Anda menggunakan model yang menghasilkan nilai probabilistik  $p(y = +1|\mathbf{x})$ . Model dari data latih mendapatkan **probabilitas** sebagai berikut untuk masing-masing contoh dalam kedua kelas yang ada:

- $y = +1$ :  $\{0.9, 0.4, 0.7, 0.8\}$
- $y = -1$ :  $\{0.1, 0.7, 0.2, 0.3, 0.2, 0.5, 0.3, 0.6\}$

Gambarkan ROC curves dengan menggunakan nilai-nilai batas (*threshold*) berikut: 0.00, 0.25, 0.45, 0.65, 1.00! (UTS Pengenalan Pola 2018)

## CONTOH (LANJUTAN)

Gambarkan ROC curves dengan menggunakan nilai-nilai batas (*threshold*) berikut: 0.00, 0.25, 0.45, 0.65, 1.00! (UTS Pengenalan Pola 2018)

- $y = +1$ :  $\{0.9, 0.4, 0.7, 0.8\}$
- $y = -1$ :  $\{0.1, 0.7, 0.2, 0.3, 0.2, 0.5, 0.3, 0.6\}$

# METRIK PADA REGRESI

- Mean Absolute Error (MAE) dan variasinya

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

- Mean Squared Error (MSE) dan variasinya

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$

Kapan kita menggunakan MSE, kapan MAE?

## METRIK LAINNYA

- **Statistical models:**  $R^2$ , Akaike information criterion (AIC), widely applicable information criterion (WAIC)
- **Information retrieval:** precision@K, mean average precision (MAP), normalized discounted cumulative gain (NDCG)
- **Text summarization:** BLEU ( $\approx$  precision), ROUGE ( $\approx$  recall)
- **Biometric:** false match rate (FMR), false non-match rate (FNMR)

Terima kasih