KOMINFO digitalent

PRO Professional Academy

Digital Talent Scholarship 2022

Structuring Machine Learning Projects

Lead a sprint through the Machine Learning Track

KOMINFO digitalent

Agenda

- Single number evaluation
- Comparing to human level performance
- Error analysis
- Mismatched training val/test set

Objektif Pembelajaran

- Memahami Orthogonalization
- Set up val and test sets
- Human level performance as a proxy for Bayes Error
- Estimate avoidable bias and variance

KOMINFO digitalent

Are your students ML-ready?

KOMINFO digitalent

Recap

ML Ideas

- Collect more data
- Collect more diverse training set
- Train algorithm longer with gradient descent
- Try Adam instead of gradient descent
- Try bigger network
- Try smaller network
- Try dropout
- Add regularization
- Network architecture

KOMINFO digitalent

ML Ideas

Kalau kita setting hyperparameter satu-satu, pasti membutuhkan waktu yang lama. Jadi kita memerlukan strategi untuk mengevaluasi model kita.

Chain of Assumptions

- Fit training set bagus pada cost function
  - Bigger network
  - Adam
- Fit val set bagus pada cost function
  - Regularization
  - Bigger training set
- Fit test set bagus pada cost function
  - Bigger val set
- Perform well in real world
  - Change val set or cost function

Single number evaluation metrics

Evaluation metric allows you to quickly tell if classifier A or classifier B is better, and therefore having a dev set plus single number evaluation metric distance to speed up iterating.

- Accuracy
- Precision
- Recall
- F1 Score

Single number evaluation metrics

Contoh kasus

$$\text{F1-Score} = \frac{2}{\frac{1}{p}+\frac{1}{r}}$$

Classifier	Precision (p)	Recall (r)	F1-Score
A	95%	90%	92.4 %
B	98%	85%	91.0%

Single number evaluation metrics

Dimana

Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot precision \cdot recall}{precision + recall}$

Single number evaluation metrics

Contoh kasus lain. Bagaimana kalau data terlalu banyak?

Algorithm	US	China	India	Other
A	3%	7%	5%	9%
B	5%	6%	5%	10%
C	2%	3%	4%	5%
D	5%	8%	7%	2%
E	4%	5%	2%	4%
F	7%	11%	8%	12%

Single number evaluation metrics

Cari rata-rata

Algorithm	US	China	India	Other	Average
A	3%	7%	5%	9%	6%
B	5%	6%	5%	10%	6.5%
C	2%	3%	4%	5%	3.5%
D	5%	8%	7%	2%	5.25%
E	4%	5%	2%	4%	3.75%
F	7%	11%	8%	12%	9.5%

Satisficing and Optimizing metrics

- **Satisficing**, yang penting terpenuhi. Meaning that it just has to be good enough.
- **Optimizing**, the class of methods that seek to maximize.

Bagaimana menentukan mana classifier terbaik?

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

Satisficing and Optimizing metrics

Kita harus tahu bagian mana yang perlu dioptimize dan bagian mana yang hanya perlu disatisfy.

Kalau satisficing metric berupa running time <=100 ms dan optimizing metric berupa Accuracy, maka **classifier B** adalah pilihan yang tepat!

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

Train/Val/Test set distribution

Syarat-syarat distribusi data:

- Harus berasal dari distribusi yang sama agar hasil maksimal.

Kalau data train berbeda dengan val dan test, ibarat kita train dengan data yang tidak cocok dengan apa yang kita inginkan. Dengan begitu pola yang dibaca tidak akan sesuai.

Coba berikan contohmu!

Train/Val/Test set distribution

Guideline

Choose a validation set and test set to reflect data you expect to get in the future and consider important to do well.

Artinya:
Pilihlah validation set dan test set yang merefleksikan data yang akan dihadapi oleh modelmu.

Size of val and test sets

Contoh kasus

- 70% train and 30% test
  Dataset kecil. 7k train dan 3k test.
- 60% train, 20% val, and 20% test
  Dataset kecil. 6k train, 2k validation, 2k test
- 98% train, 1% val, 1% test
  Dataset besar. 980k train, 10k val, 10k test

Comparing to human-level performance

- Bayes Optimal Error, most minimal error
- Mengapa setelah surpass human-level menjadi lambat?
  - Manusia mendekati bayes optimal error

Contoh kasus

- Human 1%, Train 8%, Test 10%, Focus on Bias
- Human 7.5%, Train 8%, Test 10%, Focus on Variance

Human-level error

- Normal human 3%
- Typical doctor 1%
- Experienced doctor 0.7%
- Group of experienced doctor 0.5%

Contoh kasus

Surpassing human-level performance

Masalah yang sudah surpass human-level performance

- Online Advertising
- Product recommendation
- Logistics
- Loan approvals

Note: Yang di atas adalah structured data, bukan natural perception (contoh : Computer vision, Sound recognition). Natural Perception sulit untuk mengalahkan human-level performance.

Improving Model

Avoidable bias

- Train bigger model
- Train longer/better optimization algorithms
- Better NN architecture/hyperparameter search

Variance

- More data
- Regularization
- Better NN architecture/hyperparameter search

Error Analysis

Menentukan error mana yang harus diperbaiki
Contoh kasus. Kalau ada cats and dogs dan kita tidak tahu apakah harus memperbaiki data, kita bisa mengambil 100 misrecognized data dan list jenis kesalahannya. Dan kita bisa memperbaiki masalah terbesar.

Image	Dog	Great Cats	Blurry	Comments
1	✓			Pitbull
2			✓	
3		✓	✓	Rainy day at zoo
⋮	⋮	⋮	⋮	
% of total	8%	43%	61%	

Error Analysis Table

Kalau error 10%, artinya bisa kita ambil kesimpulan bahwa kita bisa menghilangkan 61% error dari 10%, 6.1% dari total error.

Image	Dog	Great Cats	Blurry	Comments
1	✓			Pitbull
2			✓	
3		✓	✓	Rainy day at zoo
⋮	⋮	⋮	⋮	
% of total	8%	43%	61%	

Mislabeled Images

- Kalau sedikit, biarkan
- Kalau kesalahan terulang terus, bisa merusak DL
- Masukkan ke Error Analysis Table

- Cek validation dan test set
- Cek apa yang membuat salah dan apa yang membuat benar

https://youtube.com/playlist?list=PLkDaE6sCZn6E7jZ9sN_xHwSHOdjUxUW_b

[Maria Khalusova: Machine Learning Model Evaluation Metrics | PyData LA 2019 - YouTube]

KOMINFO digitalent

Q & A

KOMINFO digitalent

Thank You