Pendahuluan : Evaluasi*

Dr. rer. nat. Hendri Murfi

* Beberapa bagian dari slide ini adalah terjemahan dari slide Data Mining oleh I. H. Witten, E. Frank dan M. A. Hall

Intelligent Data Analysis (IDA) Group
Departemen Matematika, Universitas Indonesia – Depok 16424
Telp. +62-21-7862719/7863439, Fax. +62-21-7863439, Email. hendri@ui.ac.id

Evaluasi

- Seberapa prediktif model yang sudah terbentuk ? Dengan kata lain berapa kapabilitas generalisasi dari model yang terbentuk?
- Akurasi pada data training bukan merupakan indikator kinerja model untuk data di masa yang akan datang
- Jika kita memiliki cukup banyak data, solusi sederhana yang dapat digunakan adalah bagi data menjadi data training dan data testing. Selanjutnya evaluasi akurasi pada data testing.
- Akan tetapi, data biasanya terbatas, sehingga diperlukan teknik-teknik yang lebih tepat

Evaluasi

- Asumsi: baik data training dan data testing adalah sampel-sample yang representatif untuk masalah yang dihadapi
- Data testing adalah data independen yang tidak terlibat dalam pembentukan model
- Data training dan data testing boleh jadi memiliki karakteristik yang berbeda
  - Contoh: untuk mengestimasi kinerja suatu model dari kota A pada data kota lain yang berbeda, maka uji model tersebut pada data dari kota B

Evaluasi

- Setelah proses evaluasi selesai, semua data (training + testing) digunakan untuk membangun model akhir
- Secara umum, semakin besar data training maka akan semakin baik kinerja model
- Sementara itu, semakin besar data testing maka akan semakin akurat estimasi kinerja
- Machine Learning diharapkan handal untuk data training yang kecil

Isu-Isu Pada Evaluasi

- Prosedur estimasi kinerja: holdout, cross-validation, bootstrap
- Ukuran estimasi kinerja:
  - Klasifikasi: success rate, cost-sensitive, ...
  - Regresi: MSE, RMSE, MAE, RSE, RAE, ...
  - Regresi Ordinal, Ranking, Estimasi Densitas, ...
- Kehandalan estimasi kinerja: confidence interval
- Perbandingan skema pembelajaran: t-test

Prosedur Estimasi Kinerja

Metode Holdout

- Metode *holdout* membagi sejumlah data untuk testing dan menggunakan sisinya untuk training
  - Umumnya: sepertiga untuk testing, sisanya untuk training
- Masalah: sample-sample tsb boleh jadi tidak representatif
  - Contoh: suatu kelas bisa jadi tidak ada pada data testing
- Solusi: menggunakan metode *stratification*
  - Menjamin bahwa masing-masing kelas direpresentasikan dengan proporsi hampir sama pada kedua bagian data

Prosedur Estimasi Kinerja

Metode Repeated Holdout

- Metode *holdout* dapat dibuat lebih handal dengan mengulang proses untuk subsampel yang berbeda
  - Pada setiap iterasi, suatu proporsi tertentu dipilih secara acak untuk training
  - Kinerja pada setiap iterasi dirata-rata untuk mendapatkan estimasi kinerja total
- Metode ini disebut *repeated holdout*
- Masih tidak optimal: kemungkinan terjadi tumpang tindih pada data testing pada masing-masing iterasi

Prosedur Estimasi Kinerja

K-Fold Cross-Validation

- Metode *cross-validation* akan menghindari tumpang tindih pada data testing
  - Tahap 1: bagi data menjadi k bagian dengan ukuran yang sama
  - Tahap 2: gunakan masing-masing bagian untuk testing, sisanya sebagai training
- Metode seperti ini dikenal sebagai *k-fold cross validation*
- Biasanya digunakan metode *stratification* sebelum proses *cross-validation* dilaksanakan
- Estimasi tingkat kesuksesan dirata-rata untuk mendapatkan estimasi total

Prosedur Estimasi Kinerja

10-Fold Cross-Validation

- Metode standar untuk evaluasi adalah *10-fold cross-validation*
- Kenapa 10 ?
  - Banyak hasil eksperiment menunjukan bahwa ini adalah pilihan terbaik untuk mendapatkan estimasi yang akurat
  - Ada juga beberapa bukti teoritis untuk ini
- Proses *stratification* akan mereduksi variansi estimasi
- *5-fold* atau *20-fold* sering juga memberikan hasil yang hampir sama

Prosedur Estimasi Kinerja

Leave-One-Out Cross-Validation

- *Leave-One -Out* adalah bentuk khusus dari *cross-validation*, yaitu jumlah *fold* sama dengan jumlah data training
- Tidak ada proses acak dalam menentukan subsampel
- Membutuhkan biaya komputasi yang sangat mahal
- Diutamakan untuk data yang sangat kecil

Ukuran Estimasi Kinerja Klasifikasi

Success Rate

- Satuan ukuran estimasi kinerja yang umum digunakan untuk masalah klasifikasi adalah tingkat kesuksesan (*success rate*)
  - Sukses: kelas suatu data diprediksi dengan benar
  - Gagal/Error: kelas suatu data dipredikasi dengan tidak benar
  - *Success rate*: proporsi kesuksesan terhadap semua data

Ukuran Estimasi Kinerja Klasifikasi

Confusion Matrix

- *Confusion matrix*:

		Predicted class	
		Yes	**No**
Actual class	**Yes**	**True positive**	**False negative**
	No	**False positive**	**True negative**

- *True positive rate*: TP/(TP+FN)
- *False positive rate*: FP/(FP+TN)
- *Success rate*: (TP+TN)/(TP+TN+FP+FN)
- *Error rate*: 1 – *Success rate*

Ukuran Estimasi Kinerja Klasifikasi

Kappa Statistic

- Misal *confusion matrix* untuk *actual predictor* (kiri) vs. *random predictor* (kanan) adalah:

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

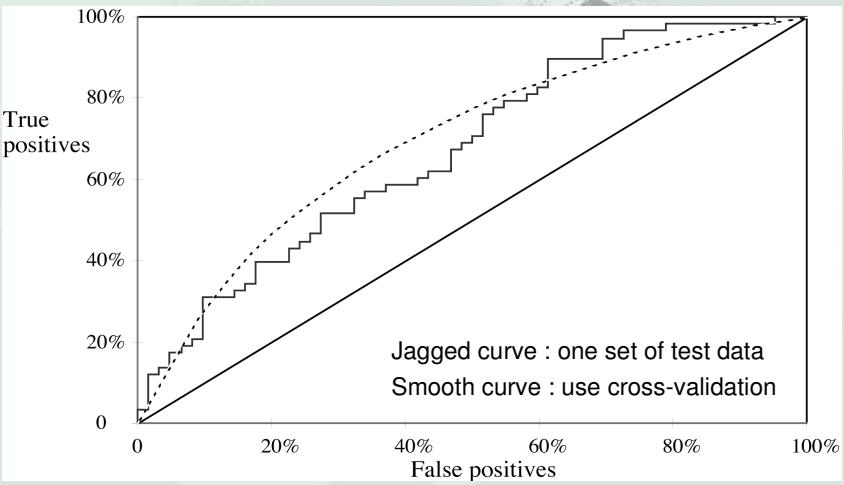
		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	60	30	10	*100*
	b	36	18	6	*60*
	c	24	12	4	*40*
	total	*120*	*60*	*20*	

- Jumlah sukses: jumlah entri pada diagonal (D)
- *Kappa statistic*: Ukuran peningkatan kinerja relatif terhadap *random predictor* → $\frac{D_{observed} - D_{random}}{D_{perfect} - D_{random}}$

Ukuran Estimasi Kinerja Klasifikasi

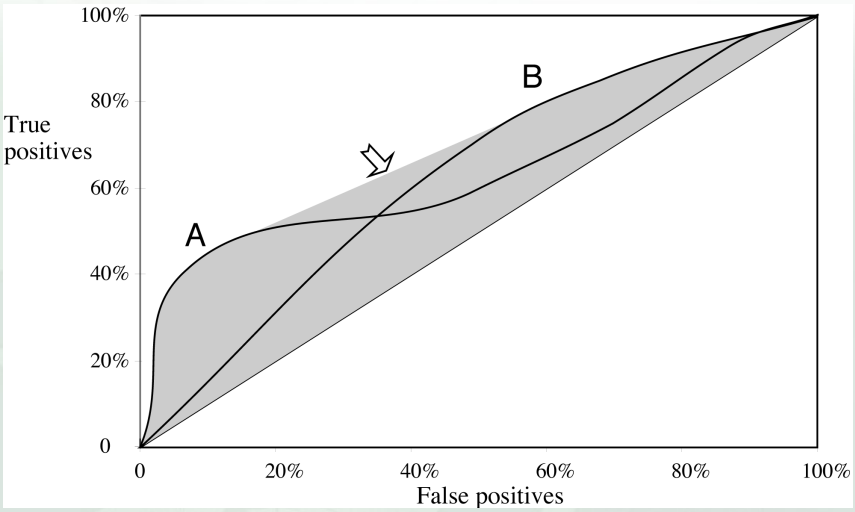
ROC Curve

- *Receiver Operating Characteristic* (ROC) *Curve* adalah suatu kurva yang mengambarkan estimasi kinerja klasifikasi untuk proporsi *false positives* vs *true positives* yang bervariasi



Ukuran Estimasi Kinerja Klasifikasi

ROC Curve



- For a small, focused sample, use method A
  - For a larger one, use method B
- In between, choose between A and B with appropriate probabilities

Ukuran Estimasi Kinerja Klasifikasi

Klasifikasi dengan Costs

- Dalam prakteknya, berbagai jenis kesalahan klasifikasi sering dikenakan biaya (*cost*) yang berbeda
- Contoh:
  - Promotional mailing
  - Terrorist profiling
  - Loan decisions
  - Fault diagnosis

Ukuran Estimasi Kinerja Klasifikasi

Cost Matrix

- Contoh dua cost matrices:

		Predicted class	
		yes	*no*
Actual class	*yes*	0	1
	no	1	0

		Predicted class		
		a	*b*	*c*
Actual class	*a*	0	1	1
	b	1	0	1
	c	1	1	0

- Success rate diganti dengan rata-rata biaya setiap prediksi
  - Cost diberikan oleh entri yang sesuai pada cost matrix

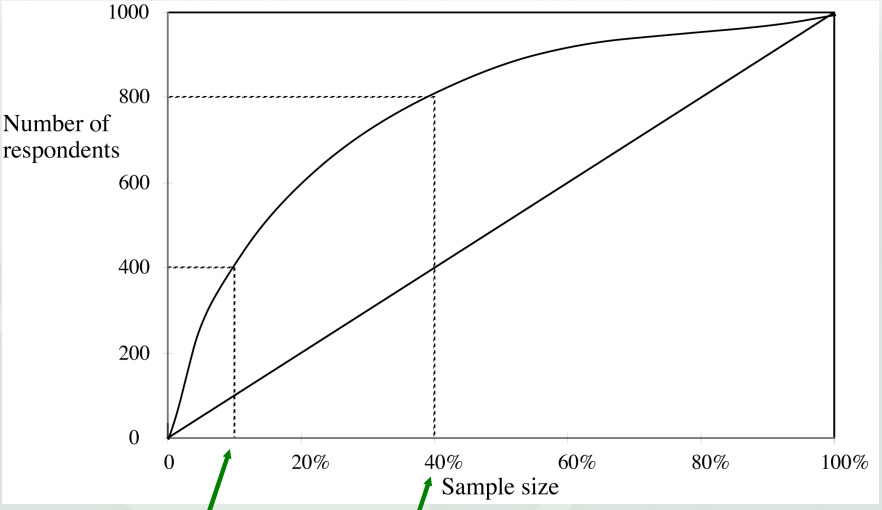
Ukuran Estimasi Kinerja Klasifikasi

Lift Charts

- Pada prakteknya, *cost* jarang diketahui
- Keputusan biasanya diambil dengan membandingkan skenario yang mungkin
- Contoh: surat promosi ke 1000000 rumah
  - Kirim ke semua (100%) → 1000 yang merespon (0.1%)
  - Kirim ke 100000 (10%) → 400 yang merespon (0.4%)
  - Kirim ke 400000 (40%) → 800 yang merespon (0.2%)
- *Lift factor* adalah faktor peningkatan respon, misal 4 pada kasus-2, dan 2 pada kasus-3. *Lift chart* memungkinkan perbadingan secara visual.

Ukuran Estimasi Kinerja Klasifikasi

Lift Charts



Ukuran Estimasi Kinerja Regresi

Mean-Squared Error

- Satuan ukuran estimasi kinerja yang populer digunakan untuk masalah regresi adalah *mean-squared error*
  - Nilai-nilai target sebenarnya: $a_1, a_2, \ldots, a_n$
  - Nilai-nilai target hasil prediksi: $p_1, p_2, \ldots, p_n$
  - *Mean-Squared Error*: $\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$

Ukuran Estimasi Kinerja Regresi

Satuan Ukuran Kinerja Lain

- *Root mean-squared error*:

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

- *Mean absolute arror* kurang sensitif terhadap outlier dibandingkan mean-squared error:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$$

Ukuran Estimasi Kinerja Regresi

Satuan Ukuran Kinerja Lain

- Kadang-kadang *relative error* lebih cocok untuk suatu keadaan, misal: 10% untuk error dari 50 ketika memprediksi 500.
- *Relative squared error*:

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(\bar{a} - a_1)^2 + \ldots + (\bar{a} - a_n)^2}$$

- *Relative absolute error*:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|\bar{a} - a_1| + \ldots + |\bar{a} - a_n|}$$

Ukuran Estimasi Kinerja Regresi

Satuan Ukuran Kinerja Lain

- *Statistical correlation* antara nilai sebenarnya dan nilai prediksi:

$$\frac{S_{PA}}{\sqrt{S_P S_A}}$$

$$S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1} \qquad S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1} \qquad S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$$

- Memiliki skala: -1 sd +1. Kinerja yang baik memiliki nilai yang besar.

Ukuran Estimasi Kinerja Regresi

Satuan Ukuran Kinerja Terbaik ?

- Terbaik adalah dengan melihat semua satuan ukuran kinerja tsb
- Contoh:

	A	B	C	D
Root mean-squared error	67.8	91.7	63.3	57.4
Mean absolute error	41.3	38.5	33.4	29.2
Root rel squared error	42.2%	57.2%	39.4%	35.8%
Relative absolute error	43.1%	40.1%	34.8%	30.4%
Correlation coefficient	0.88	0.88	0.89	0.91

- D terbaik, C terbaik kedua, A & B dapat diperdebatkan

Kehandalan Estimasi Kinerja

- Asumsikan estimasi kesuksesan adalah 75%. Seberapa dekat hasil estimasi ini terhadap tingkat kesuksesan sebenarnya ?
  - Tergantung dari jumlah data testing
- Kasus ini sangat mirip dengan pelemparan koin
  - „head“ sebagai „sukses“, „tail“ sebagai „gagal“
- Dalam statistik, kesuksesan dari kejadian-kejadian bebas seperti ini dikenal dengan nama proses Bernoulli
  - Teori statistik yang memberikan interval kepercayaan untuk hasil sebenarnya

Kehandalan Estimasi Kinerja

Proses Bernoulli

- Proses Bernoulli: tingkat kesuksesan p terletak pada suatu interval tertentu dengan tingkat kepercayaan tertentu
- Contoh 1: S = 750 sukses dalam N = 1000 percobaan
  - Estimasi tingkat kesuksesan: 75%
  - Seberapa dekat hasil ini pada tingkat kesuksesan sebenarnya p ? Jawab: dengan tingkat kepercayaan 80%, p terletak pada interval [73.2, 76.7]
- Contoh 2: S = 75 sukses dalam N = 100 percobaan
  - Estimasi tingkat kesuksesan: 75%
  - Seberapa dekat hasil ini pada tingkat kesuksesan sebenarnya p ? Jawab: dengan tingkat kepercayaan 80%, p terletak pada interval [69.1, 80.1]

Kehandalan Estimasi Kinerja

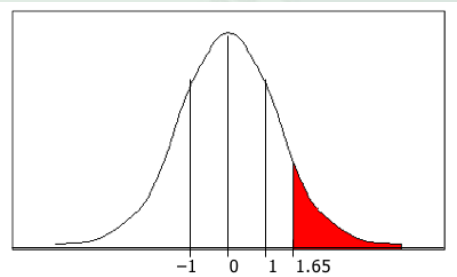
Proses Bernoulli: Interval Kepercayaan

- Probabiliti bahwa suatu variabel random X, dengan mean nol, terletak pada interval kepercayaan dengan lebar 2z adalah: $Pr[-z \le X \le z] = c$
- Untuk distribusi yang simetri: $Pr[-z \le X \le z] = 1 - 2\ Pr[x \ge z]$

Kehandalan Estimasi Kinerja

Proses Bernoulli: Interval Kepercayaan

- Interval kepercayaan untuk distribusi normal dengan mean 0 dan variansi 1 adalah:



Pr[$X \ge z$]	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

- Sehingga: $Pr[-1.65 \le X \le 1.65] = 90\%$
- Untuk penggunaannya, kita harus mereduksi variable random supaya memiliki mean 0 dan variansi 1

Kehandalan Estimasi Kinerja

Proses Bernoulli: Fakta

- Misal mean dan variansi untuk satu percobaan Bernoulli dengan *success rate* p adalah p dan p(1-p).
- Jika dilakukan N percobaan, maka ekspektasi *success rate* f = S/N adalah suatu variabel random dengan mean yang sama p, dan variansi direduksi oleh N menjadi p(1-p)/N

Kehandalan Estimasi Kinerja

Proses Bernoulli: Tranformasi ke Distribusi N(0,1)

- Untuk mentransformasikan variabel random f memiliki distribusi N(0,1), maka f dikurangi mean dan dibagi dengan standar deviasi, yaitu:

$$\frac{f-p}{\sqrt{p(1-p)/N}}$$

Sehingga:

$$Pr\left[-z \leqslant \frac{f-p}{\sqrt{p(1-p)/N}} \leqslant z\right] = c$$

- Diperoleh nilai p:

$$p = \left(f + \frac{z^2}{2N} \mp z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right)$$

Kehandalan Estimasi Kinerja

Proses Bernoulli

Contoh 1: f = 75%, N=1000, c = 80% (sehingga z = 1.28), maka: dengan tingkat kepercayaan 80%, p terletak pada interval [73.2, 76.7]

- Contoh 2: f = 75%, N=100, c = 80% (sehingga z = 1.28), maka: dengan tingkat kepercayaan 80%, p terletak pada interval [69.1, 80.1]

Catatan: asumsi distribusi normal hanya valid untuk data yang besar (N > 100). Perhatikan contoh berikut:

- Contoh 3: f = 75%, N=10, c = 80% (sehingga z = 1.28), maka: dengan tingkat kepercayaan 80%, p terletak pada interval [54.9, 88.1]