Data Mining

Pertemuan Ke 2
Feature Selection

Null hypothesis

- Pernyataan umum atau posisi default bahwa tidak ada hubungan antara dua fenomena yang diukur atau tidak ada hubungan antar kelompok

Type I dan II Error

- Type I error : penolakan Null hypothesis yang benar atau dikenal sebagai temuan atau "True Positive"
- Type II error : penerimaan Null hypothesis yang salah atau "False Negative"

VarianceThreshold

- Menghapus semua fitur yang variansnya tidak memenuhi beberapa ambang batas.
- Menghapus semua fitur varian nol, yaitu fitur yang memiliki nilai yang sama di semua sampel.

$$Var[X] = p(1-p)$$

Univariate feature selection

- Memilih fitur terbaik berdasarkan uji statistik univariat.
  - Select K Best
  - Select Percentile
  - False Positive Rate
  - False discovery rate
  - Family-wise error rate

Select K Best

- Pilih fitur sesuai dengan nilai k tertinggi.
- Jumlah k sama dengan jumlah fitur yang disisakan.

Select Percentile

- Memilih fitur menurut percentile dari skor tertinggi.
- Semakin tinggi persentase semakin banyak tolerans fitur yang tersisa

False Positive Rate

- Rasio antara jumlah peristiwa negatif yang salah dikategorikan sebagai positif (false positive) dan jumlah total peristiwa negatif aktual (regardless of classification)

$$\frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

False discovery rate

- dirancang untuk mengontrol proporsi yang diharapkan dari "penemuan" (Null Hypothesis yang ditolak) yang salah (penolakan yang salah dari hubungan Null)

False discovery rate

$$Q = V/R = V/(V+S)$$

Q : proporsi penemuan palsu di antara penemuan
V : jumlah penemuan palsu
S : jumlah penemuan benar.
R : penemuan

$$FDR = Q_e = \mathrm{E}[Q]$$

$\mathrm{E}[Q]$ nilai yang diharapkan dari Q.
Tujuannya adalah untuk menjaga FDR di bawah ambang yang diberikan q.
Untuk menghindari pembagian dengan nol, Q didefinisikan sebagai 0 jika R = 0.

Family-wise error rate

- Probabilitas dalam membuat satu atau lebih penemuan palsu, atau kesalahan tipe I saat melakukan beberapa tes hipotesis.

Multiple Hypothesis Test

	Null hypothesis Benar (H_0)	Alternative hypothesis Benar (H_A)	Total
Test dianggap significant	V : false positives (Type I error)	S : true positives (True Discovery)	R
Test dianggap tidak significant	U : true negatives	T : false negatives (Type II error)	m-R
Total	m_0	$m - m_0$	m

Family-wise error rate

$$\mathrm{FWER} = \Pr(V \geq 1)$$

Jumlah munculnya V diatas 0 dibagi semua kejadian

Atau setara dengan :

$$\mathrm{FWER} = 1 - \Pr(V = 0).$$

1 - Jumlah munculnya V = 0 dibagi semua kejadian

Mutual information

- Dua variabel acak dianggap sebagai ukuran saling ketergantungan antara dua variabel.
- Mengukur "jumlah informasi" yang diperoleh tentang satu variabel acak melalui pengamatan variabel acak lainnya.

Formula MI

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

- di mana p (x, y) adalah fungsi probabilitas gabungan dari X dan Y, dan p (x) dan p (y) adalah fungsi distribusi probabilitas marjinal masing-masing dari X dan Y.