Naïve Bayes

Ali Akbar Septiandri

November 10, 2017

untuk Astra Graphia IT

Daftar Isi

1. Naïve Bayes

2. Pros & Cons

Naïve Bayes

Klasifikasi Bayesian

- Tujuan: fungsi pembelajaran $f(x) \rightarrow y$
- Klasifikasi probabilistik: kelas yang paling mungkin jika diberikan hasil observasinya, i.e. $\hat{y} = arg\max_{y} P(y|x)$
- Probabilitas bayesian dari sebuah kelas:

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_{y'} P(x|y')P(y')}$$

Klasifikasi Bayesian: Komponen

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y'} P(x|y')P(y')}$$

- $P(y)$: probabilitas prior dari kelas
  "mana kelas yang sering muncul, mana yang jarang"
- $P(x|y)$: class-conditional model
  "seberapa sering observasi x dalam kasus y"
- $P(x)$: normalisasi

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- “Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?”

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- "Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?"
- Model generatif selalu melakukan **klasifikasi probabilistik**

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- "Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?"
- Model generatif selalu melakukan **klasifikasi probabilistik**
- Klasifikasi probabilistik tidak berarti generatif, e.g. *logistic regression*

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!
- Namun, kita mengetahui observasi untuk **masing-masing nilai** x_i untuk setiap kelas

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!
- Namun, kita mengetahui observasi untuk **masing-masing nilai** x_i untuk setiap kelas
- Asumsi yang digunakan Naïve Bayes adalah $x_1 ... x_d$ *conditionally independent* jika diberikan y

$$P(\mathbf{x}|y) = \prod_{i=1}^{d} P(x_i|x_1, ..., x_{i-1}, y) = \prod_{i=1}^{d} P(x_i|y)$$

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik
- Cuaca terik “menjelaskan” dependensi antara pergi ke pantai dan *heatstroke*

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik
- Cuaca terik “menjelaskan” dependensi antara pergi ke pantai dan *heatstroke*
- Dalam klasifikasi, nilai kelas menjelaskan hubungan antaratribut

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*:
  $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas:
  $P(setosa) = P(versicolor) = P(virginica) = 1/3$

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*:
  $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas:
  $P(setosa) = P(versicolor) = P(virginica) = 1/3$
- Asumsi: atribut terdistribusi Gaussian dan independen jika diketahui kelasnya

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas: $P(setosa) = P(versicolor) = P(virginica) = 1/3$
- Asumsi: atribut terdistribusi Gaussian dan independen jika diketahui kelasnya
- Dicocokkan dengan *maximum likelihood estimation* (MLE) untuk Gaussian, e.g.

$$\hat{\mu}_{l,setosa} = \frac{1}{50} \sum_{i;y=setosa} l_i$$

$$\hat{\sigma}^2_{l,setosa} = \frac{1}{50} \sum_{i;y=setosa} (l_i - \hat{\mu}_{l,setosa})^2$$

PDF

$$P(x|\mu,\sigma^2) = \mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

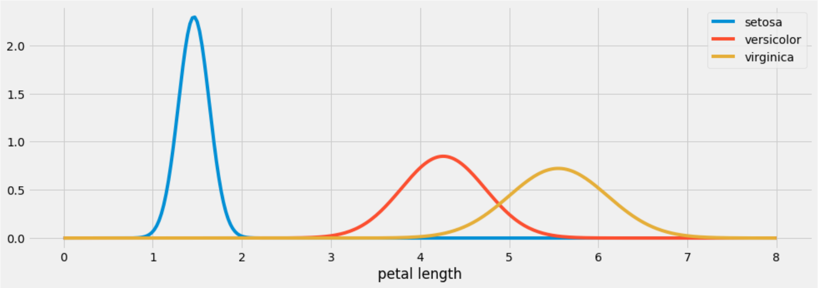
Maximum Likelihood Estimation (MLE)

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

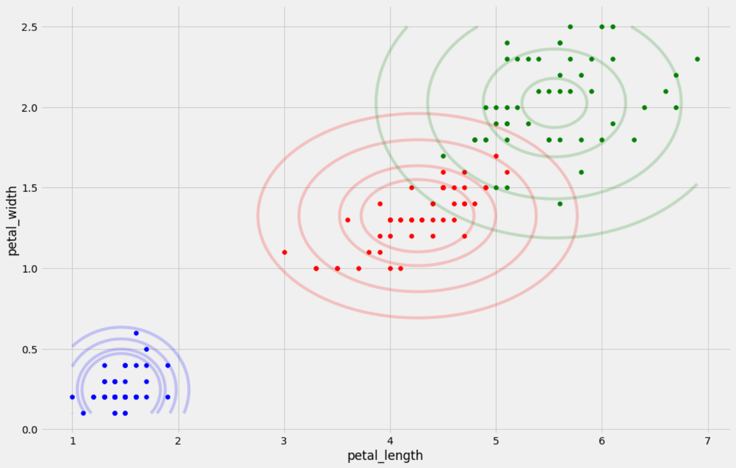
Batas Keputusan

- Beda rataan, variansi sama: garis lurus atau bidang lurus
- Rataan sama, beda variansi: lingkaran atau elips
- Kasus umum: kurva parabola



Gambar 1: Tiga Gaussians yang akan menghasilkan batas keputusan berupa garis

Batas Keputusan untuk Dua Variabel



Gambar 2: Batas keputusan dari tiga kelas

Contoh Kasus Diskrit

Asumsi: Distribusi Bernoulli

Contoh pada kasus identifikasi spam e-mail

D1: “send us your password” (s)

D2: “send us your review” (h)

D3: “review your password” (h)

D4: “review us” (s)

D5: “send your password” (s)

D6: “send us your account” (s)

Dokumen baru: “review us now”

word	**spam**	**ham**
password	2/4	1/2
review	1/4	2/2
send	3/4	1/2
us	3/4	1/2
your	3/4	1/2
account	1/4	0/2

$P(spam) = 4/6$, $P(ham) = 2/6$

Contoh Kasus Diskrit

$P(review\ \ us|spam) = P(0, 1, 0, 1, 0, 0|spam)$
$P(review\ \ us|ham) = P(0, 1, 0, 1, 0, 0|ham)$
$P(ham|review\ \ us) \approx 0.87$

Pros & Cons

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

- Nilai ϵ contohnya 1 atau 0.5, tetapi bisa juga dengan $num(w)/num$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

- Nilai ϵ contohnya 1 atau 0.5, tetapi bisa juga dengan $num(w)/num$
- Kasus ini sering terjadi karena Zipf's law (50% kata hanya muncul sekali)

Masalah Conditional Independence

- Asumsi ini pada banyak kasus kurang tepat, terlalu naif
- Setiap kasus dianggap berkontribusi sama kepada kelas
- *Classifier* yang dihasilkan dapat ditipu dengan memperbanyak kata-kata yang mengindikasikan bahwa e-mail tersebut “ham”

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i =?, ..., X_d = x_d | y)$?

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i = ?, ..., X_d = x_d | y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i = ?, ..., X_d = x_d | y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*
- Hitung saja berdasarkan atribut yang bernilai!

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i = ?, ..., X_d = x_d | y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*
- Hitung saja berdasarkan atribut yang bernilai!
- Nilai yang hilang tersebut tidak perlu diganti

Incremental Updates

- Dengan menyimpan data dalam bentuk jumlah, data baru dapat dimutakhirkan dengan menambahkannya ke variabel yang sudah ada
- Saat perlu diklasifikasi, baru hitung nilai yang dibutuhkan
- Berlaku untuk kasus diskrit maupun kontinu

Salindia ini dipersiapkan dengan sangat dipengaruhi oleh:
Victor Lavrenko (2014)

Terima kasih