Machine Learning

Section 6: The Gaussian distribution

第6节：高斯分布

Stefan Harmeling

13. October 2021

What happend so far:

- Probability theory as an extension of propositional logic.
- Probability theory for discrete and continuous variables.
- Graphical models as a representation for PDFs with conditional independences.

概率论作为命题逻辑的延伸。
离散变量和连续变量的概率理论。
图形模型作为具有条件独立性的PDF的表示方法

What is inference?

考虑二进制变量，它们要么是真，要么是假。
逻辑推理

Consider binary variables which are either true or false.

- Logical reasoning

- define axioms定义公理
- use inference rules to deductively derive new facts 使用推理规则来推导新的事实
- can only say something about true and false 只能说一些真和假的事情
- monotonic reasoning: more knowledge makes more stuff true, never单调推理

  turns a statement “back” to false (eg. “penguins are birds”)更多的知识使更多的东西成为真的，永远不会使一个陈述 "回到 "假的状态

- Probabilistic reasoning概率论推理

- define joint probability distribution, e.g. $p(X, Y, Z|H)$定义联合概率分布
- condition on the known facts, e.g. $Z = z$*以已知事实为条件*

$$p(X, Y|Z = z, H) = p(X, Y, Z = z|H)/p(Z = z|H)$$

  called *conditioning* (aka product rule)

- integrate out the non-interesting random variables, e.g. Y

$$p(X|z, H) = \int p(X, Y|Z = z, H)dy$$

  called *marginalization* (aka sum rule) 边缘化

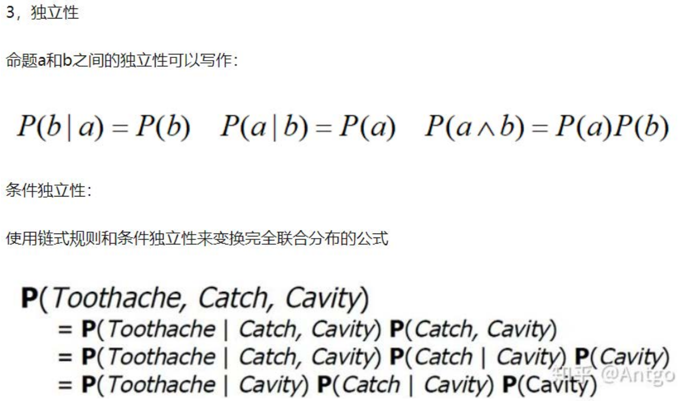
- get posterior probability of X assuming $Z = z$, i.e. $P(X|z, H)$得到假设的X的后验概率

1，先验概率

先验概率又称无条件概率，即不需要推理就知道的概率

2，后验概率

后验概率又称条件概率，根据某个已经发生的事情来推测另个事件的概率。



3，独立性

命题a和b之间的独立性可以写作：

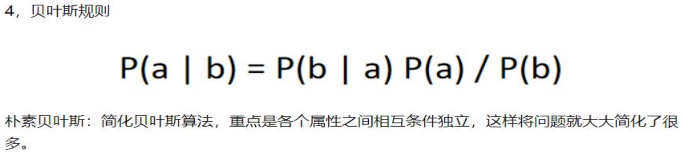
$$P(b\,|\,a) = P(b) \quad P(a\,|\,b) = P(a) \quad P(a \wedge b) = P(a)P(b)$$

条件独立性：

使用链式规则和条件独立性来变换完全联合分布的公式

$$\begin{aligned}\mathbf{P}(\mathit{Toothache, Catch, Cavity}) &\\ = \mathbf{P}(\mathit{Toothache}\,|\,\mathit{Catch, Cavity})\,\mathbf{P}(\mathit{Catch, Cavity}) &\\ = \mathbf{P}(\mathit{Toothache}\,|\,\mathit{Catch, Cavity})\,\mathbf{P}(\mathit{Catch}\,|\,\mathit{Cavity})\,\mathbf{P}(\mathit{Cavity}) &\\ = \mathbf{P}(\mathit{Toothache}\,|\,\mathit{Cavity})\,\mathbf{P}(\mathit{Catch}\,|\,\mathit{Cavity})\,\mathbf{P}(\mathit{Cavity}) &\end{aligned}$$

知乎 @Antgo



4，贝叶斯规则

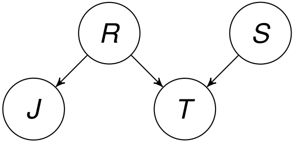
$$P(a\ |\ b) = P(b\ |\ a)\ P(a)\ /\ P(b)$$

朴素贝叶斯：简化贝叶斯算法，重点是各个属性之间相互条件独立，这样将问题就大大简化了很多。

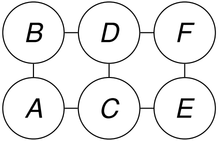
Graphical models图形化模型

. . . efficiently represent probability distributions with many variables. ......有效地表示有许多变量的概率分布

- Directed graphical models (e.g. Bayes nets)



- Undirected graphical models (not part of this lecture)



- Probabilistic programming (not part of this lecture)

```
t[0] = coin(0.4); i = 0;
while t[i] is HEAD, t[i++] = coin(0.4);
```

A few technical terms 几个技术术语

- Bayes rule

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

- unkown x (often some parameter), known y (typically the data)

未知的X（通常是一些参数）， 已知的Y（通常是数据）

- prior $p(x)$, “my belief about x before seeing data”

先验p(x)，"在看到数据之前我对x的信念"

- likelihood $p(y|x)$, “how likely is the data y for fixed value of x”, we可能性p(y|x)，"对于固定的x值，数据y的可能性有多大"，我们 say “likely” because $p(y|x)$ as a function of x is not a probability distribution since it is not normalized不是一个概率分布，因为它没有被规范化

- evidence $p(y)$, usually calculated as the integral of the nominator, renormalizes the joint $p(x, y)$证据p(y)，通常作为提名者的积分计算。将联合p(x, y)重新规范化。

- posterior $p(x|y)$, “what do I know about x after seeing data”

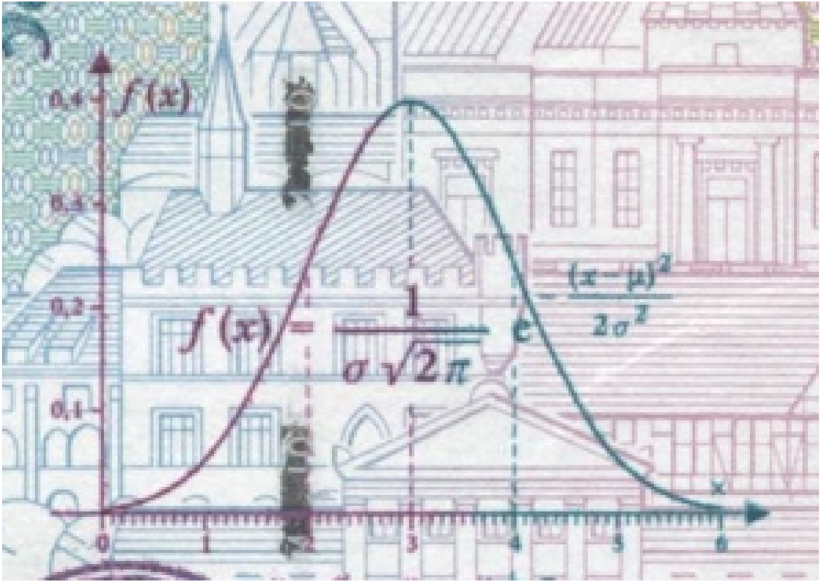
看到数据后，我对X有什么了解"

- $p(y|x)$ as a function of y for fixed x is a probability, as a function of x for fixed y it is a likelihood (confusing...)对固定的y，x的可能性是一个混乱的

- probabilities are normalized, likelihoods are not概率是标准化的，似然不是

A famous distribution一个著名的分布

GU5672972S2
DEUTSCHE BUNDESBANK
Banknote
f(x)
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
10
1777–1855 Carl Friedr. Gauß
10
ZEHN DEUTSCHE MARK
GU5672972S2
Deutsche Bundesbank
Frankfurt am Main
1. September 1999


0,4
$f(x)$
0,3
0,2
0,1
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
0
1
2
3
4
5
6
x

Definition 6.1 (Univariate Gaussian distribution)单变量正态分布

The PDF of an univariate Gaussian RV X is

$$p(x) = \mathcal{N}(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with x, μ, σ^2 being scalars, and $\pi = 3.14159265\ldots$

Notes

- μ is the mean of X, since 是X的平均值

$$\mu = \int x \mathcal{N}(x, \mu, \sigma^2)\, dx = \mathrm{E}_x x$$

- σ^2 is the variance of X *σ 2是X的方差*

$$\sigma^2 = \int (x-\mu)^2 \mathcal{N}(x, \mu, \sigma^2)\, dx = \mathrm{E}_x (x-\mu)^2$$

- These two integrals are not trivial! E.g. look at `https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-va` for a detailed derivation.
- σ is called the *standard deviation* of X *σ 被称为X的标准差*
- why write σ^2? this ensures positivity of the variance.

Lemma 6.2

1. *$\mathcal{N}$ is probability density function:* N是概率密度函数

$$\mathcal{N}(x, \mu, \sigma^2) \geq 0 \qquad \int \mathcal{N}(x, \mu, \sigma^2)\, dx = 1$$

2. *Symmetry in x and* μ在x和μ中的对称性

$$\mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{x}, \sigma^2)$$

3. *Exponential of a second degree polynomial*二度多项式的指数

$$\mathcal{N}(x, \mu, \sigma^2) = \exp(a + \eta x - \frac{1}{2}\lambda^2 x^2)$$

with $\eta = \sigma^{-2}\mu$, $\lambda^2 = \sigma^{-2}$, $a = -\frac{1}{2}\left(\log(2\pi) - \log\lambda^2 + \lambda^{-2}\eta^2\right)$.
η and λ^2 (aka precision) are called canonical or natural parameters. η 和 λ 2（又称精度）被称为典范或自然参数

4. *Any second degree polynomial $a + bx - 0.5cx^2$ with $c > 0$ induces*任何二度多项式，都会诱导出 *an (unnormalized) Gaussian distribution via $\eta = b$ and $\lambda^2 = c$, that can be normalized by adjusting a.*可以通过调整a来归一化

Inference with univariate Gaussians用单变量高斯进行推理

Assume

$$p(x) = \mathcal{N}(x, \mu, \sigma^2) = \exp(a + bx - 0.5cx^2) \qquad \text{prior 之前}$$

$$p(y|x) = \mathcal{N}(y, x, \tau^2) = \exp(d + ex - 0.5fx^2) \qquad \text{likelihood}$$

Assume μ, σ^2, τ^2 fixed and known.

What is the posterior $p(x|y)$?

$$\begin{aligned} p(x|y) &= \frac{p(y|x)\,p(x)}{\int p(y|x)\,p(x)\,dx} \\ &= \frac{\mathcal{N}(x, y, \tau^2)\,\mathcal{N}(x, \mu, \sigma^2)}{\int p(y|x)\,p(x)\,dx} \\ &= \frac{\exp\left((a + d) + (b + e)x - 0.5(c + f)x^2\right)}{\int p(y|x)\,p(x)\,dx} \\ &= \mathcal{N}(x, \nu, \xi^2) = \mathcal{N}\left(x, \frac{\sigma^{-2}\mu + \tau^{-2}y}{\sigma^{-2} + \tau^{-2}}, \frac{1}{\sigma^{-2} + \tau^{-2}}\right) \end{aligned}$$

Notes

- With

$$b = \sigma^{-2}\mu \qquad c = \sigma^{-2}$$

$$e = \tau^{-2}y \qquad f = \tau^{-2}$$

we get for the posterior variance $\xi^2 = (c + f)^{-1} = \frac{1}{\sigma^{-2}+\tau^{-2}}$ and for the posterior mean $\nu = \xi^{-2}(b + e) = \frac{\sigma^{-2}\mu+\tau^{-2}y}{\sigma^{-2}+\tau^{-2}}$

- We don’t have to calculate the normalization since we know it is我们不需要计算归一化，因为我们知道它是二阶多项式的指数，所以它将被适当地归一化。 the exponential of a second order polynomial, so it will be properly normalizable.
- The denominator does not depend on x, since x is integrated out.

分母不依赖于x，因为x被整合掉了。后验平均数是 μ 和y的加权平均数。

- The posterior mean is the weighted average of μ and y.
- So, a Gaussian prior is a *conjugate prior* for a Gaussian likelihood. 所以，高斯先验是高斯似然的共轭先验。

Definition 6.3 (Multivariate Gaussian distribution)多变量高斯分布

The PDF of an multivariate Gaussian RV X is

$$p(x) = \mathcal{N}(x, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

with x, μ being n-vectors, Σ being a symmetric positive definite $n \times n$-matrix *x, u是n个向量，Σ是一个对称的正定n×n矩阵*

Notes

- μ is the mean of X, since $\mathrm{E}_x\, x = \mu$. 是X的平均值
- Σ is the covariance of X, since $\mathrm{E}_x(x-\mu)(x-\mu)^{\mathrm{T}} = \Sigma$. Σ 是X的协方差
- $|\Sigma|$ is the determinant of Σ 是Σ的行列式
- The matrix A is positive definite iff all eigenvalues are positive.矩阵A是正定的，当且仅当所有的特征值都是正的

This corresponds to positivity for scalars.

- The univariate case is a special case of the multivariate one with单变量情况是多变量情况的一个特例，有 $n = 1$ and $\Sigma = \sigma^2$. ▸ $\delta(x, \mu) = (x-\mu)^T\Sigma^{-1}(x-\mu)$ is called Mahalanobis distance (having elliptical isolines).

Lemma 6.4

1. $\mathcal{N}$ *is probability density function:*1. N是概率密度函数。

$$\mathcal{N}(x, \mu, \Sigma) \geq 0 \qquad \int \mathcal{N}(x, \mu, \Sigma)\, dx = 1$$

2. *Symmetry in x and μ* . 在x和μ中的对称性

$$\mathcal{N}(x, \mu, \Sigma) = \mathcal{N}(\mu, x, \Sigma)$$

3. *Exponential of a second degree polynomialv* 二度多项式的指数

$$\mathcal{N}(x, \mu, \Sigma) = \exp(a + \eta^T x - \frac{1}{2} x^T \Lambda x)$$

with $\eta = \Sigma^{-1}\mu$, $\Lambda = \Sigma^{-1}$, $a = -\frac{1}{2}\left(n \log(2\pi) - \log|\Lambda| + \eta^T \Lambda^{-1} \eta\right)$.
Parameters η *and* Λ *(aka precision matrix) are called canonical or natural.*参数 η 和Λ （又称精度矩阵）被称为典范或自然

4. *Any second degree polynomial* $a + b^T x - 0.5 x^T C x$ *with* C *positive definite induces an (unnormalized) Gaussian distribution via* $\eta = b$ *and* $\Lambda = C$*, that can be normalized by adjusting a.*定义通过 η =b和Λ =C诱导出一个（未归一化的）高斯分布，可以通过调整a来归一化

Inference with multivariate Gaussians

Assume

$$
\begin{aligned}
p(x) &= \mathcal{N}(x, \mu, \Sigma) = \exp(a + b^T x - 0.5 x^T C x) && \text{prior} \\
p(y|x) &= \mathcal{N}(y, x, T) = \exp(d + e^T x - 0.5 x^T F x) && \text{likelihood}
\end{aligned}
$$

Assume μ, Σ, T fixed and known.

What is the posterior $p(x|y)$?

$$
\begin{aligned}
p(x|y) &= \frac{p(y|x)\, p(x)}{\int p(y|x)\, p(x)\, dx} \\
&= \frac{\mathcal{N}(x, y, T)\, \mathcal{N}(x, \mu, \Sigma)}{\int p(y|x)\, p(x)\, dx} \\
&= \frac{\exp\left((a+d) + (b+e)^T x - 0.5 x^T (C+F) x\right)}{\int p(y|x)\, p(x)\, dx} \\
&= \mathcal{N}(x, \nu, \Xi) = \mathcal{N}\left(x, (\Sigma^{-1} + T^{-1})^{-1}(\Sigma^{-1}\mu + T^{-1} y), (\Sigma^{-1} + T^{-1})^{-1}\right)
\end{aligned}
$$

Notes

- With

$$b = \Sigma^{-1}\mu \qquad C = \Sigma^{-1}$$
$$e = T^{-1}y \qquad F = T^{-1}$$

  we get for the posterior variance $\Xi = (C + F)^{-1} = (\Sigma^{-1} + T^{-1})^{-1}$ and for the posterior mean $\nu = \Xi^{-1}(b + e) = (\Sigma^{-1} + T^{-1})^{-1}(\Sigma^{-1}\mu + T^{-1}y)$
- We don't have to calculate the normalization since we know it is the exponential of a second order polynomial, so it will be properly normalizable.
- The denominator does not depend on x, since x is integrated out.
- The posterior mean is the weighted average of μ and y.
- So, a Gaussian prior is a *conjugate prior* for a Gaussian likelihood.

Lemma 6.5 (Gaussian joints)

A Gaussian prior and likelihood高斯先验和似然

$$p(x) = \mathcal{N}(x, \mu, \Sigma)$$
$$p(y|x) = \mathcal{N}(y, x, T)$$

induce a Gaussian joint distribution诱导出高斯联合分布

$$\begin{aligned} p(x, y) &= p(y|x)\, p(x) \\ &= \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma^{-1} + T^{-1} & -T^{-1} \\ -T^{-1} & T^{-1} \end{bmatrix}^{-1}\right) \end{aligned}$$

and Gaussian evidence

$$p(y) = \mathcal{N}(y, \mu, T + \Sigma)$$

Lemma 6.6 (Product rule for Gaussians)

The product rule $p(y|x)\,p(x) = p(x|y)\,p(y)$ reads for Gaussians

$$\mathcal{N}(y, x, T)\,\mathcal{N}(x, \mu, \Sigma) = \mathcal{N}(x, \nu, \Xi)\,\mathcal{N}(y, \mu, T + \Sigma)$$

with

$$\begin{aligned} \Xi &= (\Sigma^{-1} + T^{-1})^{-1} \\ \nu &= \Xi(\Sigma^{-1}\mu + T^{-1}y) \end{aligned}$$

where ν depends on y, but μ, T, Σ does not depend on x.

Alternatively we can write

$$\mathcal{N}(x, a, A)\,\mathcal{N}(x, b, B) = \mathcal{N}(x, c, C)\,\mathcal{N}(a, b, A + B)$$

with

$$\begin{aligned} C &= (A^{-1} + B^{-1})^{-1} \\ c &= C(A^{-1}a + B^{-1}b) \end{aligned}$$

Lemma 6.7 (Gaussian marginals and conditionals)边缘分布

A Gaussian joint distribution

$$p(x,y) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

has Gaussian marginals

$$p(x) = \int p(x,y)\,dy = \mathcal{N}(x,\mu,A)$$
$$p(y) = \int p(x,y)\,dx = \mathcal{N}(y,\nu,C)$$

and Gaussian conditionals

$$p(x|y) = p(x,y)/p(y) = \mathcal{N}(x,\mu + BC^{-1}(y-\nu), A - BC^{-1}B^T)$$
$$p(y|x) = p(x,y)/p(x) = \mathcal{N}(y,\nu + B^TA^{-1}(x-\mu), C - B^TA^{-1}B)$$

Lemma 6.8 (Sum rule for Gaussians)

The sum rule $p(y) = \int p(x,y)\,dx$ *reads for Gaussians*

$$\mathcal{N}(y,\nu,C) = \int \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) dx$$

similar for $p(x) = \int p(x,y)\,dy$

$$\mathcal{N}(x,\mu,A) = \int \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) dy$$

Lemma 6.9 (Linear transformation of a Gaussian)线性变换)

Assume random variable x is Gaussian distributed, i.e.

$$p(x) = \mathcal{N}(x, \mu, \Sigma)$$

Then any linear transformation $y = Ax + b$ of x (with matrix A and vector b) is also Gaussian distributed as follows:

$$p(y) = \mathcal{N}(y, A\mu + b, A\Sigma A^T)$$

Thus the sum $z = x + y$ of two independent Gaussian random variables x and y is also Gaussian, because

$$z = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

By the way, the convolution (“Faltung”) of two Gaussian PDFs is a Gaussian PDF.卷积

More notation!

到现在为止，我们写表示高斯PDF p(x)是x的函数，它的平均值μ和它的协方差矩阵Σ。注意，小x是大写字母的RV X的可能值。

Until now we wrote

$$p(x) = \mathcal{N}(x, \mu, \Sigma)$$

to denote that the Gaussian PDF p(x) is a function of x, its mean μ and its covariance matrix Σ. Note that small x is a possible value of the RV X with a capital letter.

Sometimes we write also

$$p(x) = p(x|\mu, \Sigma) = \mathcal{N}(x|\mu, \Sigma)$$

to directly specify the distribution of X even stressing the fact that the mean and covariance can be seen as random variables themselves.

有时我们也写直接指定X的分布，甚至强调平均数和协方差可以被看作是随机变量本身的事实。

Summary of rules for the Gaussian

1. products of Gaussians are Gaussians
2. marginals of Gaussians are Gaussians
3. conditionals of Gaussians are Gaussians
4. affine linear mappings of Gaussians are Gaussians

   Gaussian are for probability theory what affine linear mappings are for algebra. [This is a deep insight, I got from Philipp Hennig.]

Notes

- Both are represented with a matrix and a vector.
- Both are used to approximate more complicated stuff (Laplace's method/approximation vs. linear approximation).
- More work is required to clarify the exact relationship.