Machine Learning

Section 6: The Gaussian distribution

Stefan Harmeling

25. 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.

What is inference?

Consider binary variables which are either true or false.

- Logical reasoning
  - define axioms
  - use inference rules to deductively derive new facts
  - can only say something about true and false
  - monotonic reasoning: more knowledge makes more stuff true, never turns a statement “back” to false (eg. “penguins are birds”)
- Probabilistic reasoning
  - define joint probability distribution, e.g. $p(X, Y, Z|H)$
  - condition on the known facts, e.g. $Z = z$

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

    called *conditioning* (aka product rule)
  - integrate out the non-interesting random variables, e.g. Y

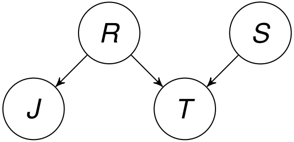
    $$p(X|z, H) = \int p(X, Y|Z = z, H)dy$$

    called *marginalization* (aka sum rule)
  - get posterior probability of X assuming $Z = z$, i.e. $P(X|z, H)$

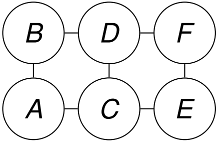
Graphical models

…efficiently represent probability distributions with many variables

- 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

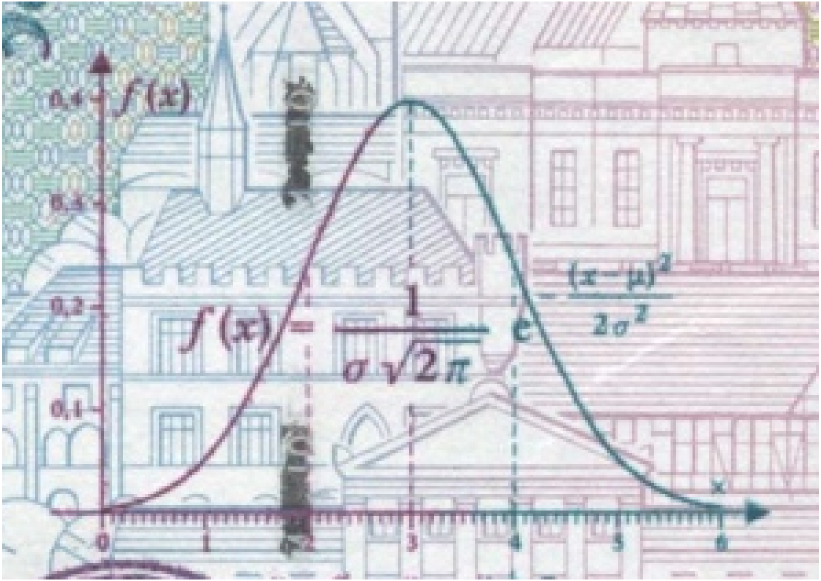
- Bayes rule

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

- unkown x (often some parameter), known y (typically the data)
- prior $p(x)$, "my belief about x before seeing data"
- likelihood $p(y|x)$, "how likely is the data y for fixed value of x", we say "likely" because $p(y|x)$ as a function of x is not a probability distribution since it is not normalized
- evidence $p(y)$, usually calculated as the integral of the nominator, renormalizes the joint $p(x, y)$
- posterior $p(x|y)$, "what do I know about x after seeing data"
- $p(y|x)$ as a function of y for fixed x is a probability, as a function of x for fixed y it is a likelihood (confusing...)
- probabilities are normalized, likelihoods are not

A famous distribution


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


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

Definition 6.1 (Univariate Gaussian distribution)

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

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

- σ^2 is the variance of X

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

- 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
- why write σ^2? this ensures positivity of the variance.

Lemma 6.2

1. *$\mathcal{N}$ is probability density function:*

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

2. *Symmetry in x and μ*

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

3. *Exponential of a second degree polynomial*

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

   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.

4. *Any second degree polynomial $a + bx - 0.5cx^2$ with $c > 0$ induces an (unnormalized) Gaussian distribution via $\eta = b$ and $\lambda^2 = c$, that can be normalized by adjusting a.*

Inference with univariate Gaussians

Assume

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

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

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 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.

Definition 6.3 (Multivariate Gaussian distribution)

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

Notes

- μ is the mean of X, since $\mathrm{E}_x\, x = \mu$.
- Σ is the covariance of X, since $\mathrm{E}_x (x - \mu)(x - \mu)^\mathrm{T} = \Sigma$.
- $|\Sigma|$ is the determinant of Σ
- The matrix A is positive definite iff all eigenvalues are positive. This corresponds to positivity for scalars.
- The univariate case is a special case of the multivariate one with $n = 1$ 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:*

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

2. *Symmetry in x and μ*

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

3. *Exponential of a second degree polynomial*

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

   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.

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.*

Inference with multivariate Gaussians

Assume

$$p(x) = \mathcal{N}(x, \mu, \Sigma) = \exp(a + b^T x - 0.5 x^T C x) \qquad \text{prior}$$

$$p(y|x) = \mathcal{N}(y, x, T) = \exp(d + e^T x - 0.5 x^T F x) \qquad \text{likelihood}$$

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

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

induce a Gaussian joint distribution

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

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

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

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

$$C = (A^{-1} + B^{-1})^{-1}$$
$$c = C(A^{-1}a + B^{-1}b)$$

Lemma 6.7 (Gaussian marginals and conditionals)

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

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

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)

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.

More notation!

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.

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.