INF367A: Probabilistic machine learning

Lecture 6: Bayesian modeling II

Pekka Parviainen

University of Bergen

11.2.2020

Outline

Recap: Bayes theorem

- We want the distribution of the parameters given the observed data:

$$P(\text{model} \mid \text{data})$$

- We can use the Bayes theorem:

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})P(\text{model})}{P(\text{data})}$$

- $P(\text{model} \mid \text{data})$: Posterior probability of parameters after observing data
- $P(\text{data} \mid \text{model})$: Likelihood
- $P(\text{model})$: Prior probability of parameters before observing data
- $P(\text{data})$: Normalizing constant

General recipe for Bayesian inference

- θ = Things we want to know
- D = Things we know
- We always need to specify the likelihood $P(D \,|\, \theta)$ and the prior $P(\theta)$!
  - The likelihood depends on the data generating process
  - The prior is your subjective belief about the quantity of interest
- Then, we compute

$$
\begin{aligned}
P(\theta \,|\, D) &= \frac{P(D \,|\, \theta) P(\theta)}{\int_\theta P(D \,|\, \theta) P(\theta) \mathrm{d}\theta} \\
&\propto P(D \,|\, \theta) P(\theta)
\end{aligned}
$$

Even more general recipe for Bayesian inference

- θ = Things we do not know but would want to know
- Z = Thing we do not know and do not care
- D = Things we know
- Marginalize out all the variables you are not interested in:

$$P(\theta \,|\, D) = \frac{\int_Z P(D \,|\, \theta, Z) P(\theta, Z) \mathrm{d}Z}{\int_\theta \int_Z P(D \,|\, \theta, Z) P(\theta, Z) \mathrm{d}Z \mathrm{d}\theta}$$

- Note: If you have discrete variables, just replace integration with summation

Operations you need to know

- Factorization (Chain rule):

$$P(A_1, A_2, \ldots, A_k) = \prod_{i=1}^{k} P(A_i \mid A_{i+1}, A_{i+2}, \ldots, A_k).$$

  - Conditional independencies can simplify the conditional distributions!
- Marginalization
  - Continuous variables:

$$P(A) = \int_B P(A, B)\mathrm{d}B$$

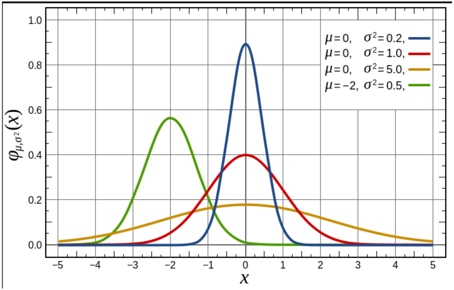
  - Discrete variables:

$$P(A) = \sum_B P(A, B)$$

Gaussian distribution

- $X \sim N(\mu, \sigma^2)$
- Parameters: μ: mean, σ^2: variance
- Inverse of the variance, $\lambda = 1/\sigma^2$, is called the precision
- Standard deviation σ
- 95% credible interval equals approximately $[\mu - 2\sigma, \mu + 2\sigma]$
- PDF:

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



Gaussian (or normal) distribution (wikip.)

Estimation of the mean of a Gaussian

- Assume that data points are sampled independently from a Gaussian with unknown mean μ and known variance σ^2 (precision $\lambda = 1/\sigma^2$)
- Data: $D = \{x_i\}_{i=1}^n$ (n independent data points)
- Likelihood:

$$P(D \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

ML estimation

- ML estimate for the mean:

$$\mu_{ML} = \arg\max_{\mu} P(D \,|\, \mu)$$

- Logarithm is a monotone function and therefore $\arg\max_{\mu} P(D \,|\, \mu) = \arg\max_{\mu} \log P(D \,|\, \mu)$
- Solve

$$\frac{\partial \log P(D \,|\, \mu)}{\partial \mu} = \frac{-2n(\bar{x} - \mu)}{2\sigma^2} = 0,$$

  where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- ML estimate:

$$\mu_{ML} = \bar{x}$$

Bayesian estimation of the mean of a Gaussian (1/2)

- Suppose we have observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \sigma^2)$, where σ^2 is known.
- To learn μ, we specify a prior

$$\mu \sim N(\mu_0, \tau_0^2)$$

- Posterior

$$\begin{aligned}
P(\mu|x) &= \frac{P(x|\mu)P(\mu)}{P(x)} \propto P(\mu)P(x|\mu) \\
&= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\
&\propto e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \\
&= \ldots
\end{aligned}$$

Bayesian estimation of the mean of a Gaussian (2/2)

- After some manipulations, we end up with the posterior

$$
\begin{aligned}
P(\mu|x) &\propto e^{-\frac{1}{2\tau_n^2}(\mu-\mu_n)^2} \\
&\propto N(\mu|\mu_n, \tau_n^2)
\end{aligned}
$$

where

$$
\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{x}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}.
$$

- Posterior precision $1/\tau_n^2$: sum of prior precision $1/\tau_0^2$ and data precision n/σ^2
- Posterior mean μ_n: precision weighted average of prior mean μ_0 and data mean $\overline{x}$.

Conjugate prior distributions (1/2)

- In the previous example

$$\text{Prior: } \mu \sim N(\mu_0, \tau_0^2)$$
$$\text{Posterior: } \mu \sim N(\mu_n, \tau_n^2).$$

  If the prior and posterior belong to the same family of distributions, we say that the prior is conjugate to the likelihood used.

  - For example, normal prior $\mu \sim N(\mu_0, \tau_0^2)$ is conjugate to the normal likelihood $N(x|\mu, \sigma^2)$.

- Conjugacy is useful, because it makes computations easy.

Conjugate prior distributions (2/2)

- With conjugate prior, the posterior is available in a closed form from

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

  - Drop all terms not depending on θ
  - Recognize the result as a density function belonging to the same family of distributions as the prior $P(\theta)$, but with a different parameter θ.

- Examples (likelihood - conjugate prior):
  - Likelihood for normal mean - Normal prior
  - Likelihood for normal variance - Inverse-Gamma prior
  - Bernoulli - Beta
  - Binomial - Beta
  - Exponential - Gamma
  - Poisson - Gamma

Exponential family*

- A distribution belongs to the exponential family if the density function can be written in form

$$f(x \mid \theta) = h(x)g(\theta)e^{\eta(\theta)\cdot T(x)}$$

- All exponential family distributions (used as likelihood) have a conjugate prior

Conjugate prior example (1/2)

- Suppose we have observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \lambda^{-1})$, where μ is known.
- To learn the precision λ, we specify a prior

$$\lambda \sim \mathrm{Gam}(a, b)$$

Gamma distribution

- Distribution for positive real values.

$$\lambda \sim \text{Gam}(a, b), \quad a > 0 : \textbf{shape}, \quad b > 0 : \textbf{rate}$$

$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$$

- Alternative parameterization uses $\lambda \sim$Gam(a, θ), $\theta = 1/b$ is called the **scale**

$$\text{Gam}(\lambda|a, \theta) = \frac{1}{\Gamma(a)\theta^a} \lambda^{a-1} e^{-\lambda/\theta}$$

Conjugate prior example (2/2)

- Observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \lambda^{-1})$, where μ is known; $\lambda \sim$Gam(a, b).

$$
\begin{aligned}
P(\lambda|x) &\propto P(x|\lambda)P(\lambda) \\
&= \prod_{i=1}^{n} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(x_i-\mu)^2} \times \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda} \\
&\propto \lambda^{\frac{n}{2}} e^{-\frac{\lambda}{2}\sum_i (x_i-\mu)^2} \times \lambda^{a-1} e^{-b\lambda} \\
&= \lambda^{\frac{n}{2}+a-1} e^{-\lambda\left[\frac{1}{2}\sum_i (x_i-\mu)^2 + b\right]} \\
&\propto \mathrm{Gam}(\lambda|a_n, b_n),
\end{aligned}
$$

with

$$
\begin{aligned}
a_n &= a + \frac{n}{2} \\
b_n &= b + \frac{1}{2}\sum_i (x_i - \mu)^2
\end{aligned}
$$

Gaussian distribution, unknown mean and precision (1/2)

- Suppose we have observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \lambda^{-1})$, where both the mean μ and the precision λ are unknown.
- The conjugate prior distribution is the normal-gamma distribution

$$\begin{aligned} P(\mu, \lambda | \mu_0, \beta, a, b) &= N(\mu | \mu_0, (\beta\lambda)^{-1}) \mathrm{Gam}(\lambda | a, b) \\ &\equiv \text{Normal-Gamma}(\mu, \lambda | \mu_0, \beta, a, b) \end{aligned}$$

Note the dependency of the prior of μ on the value of λ.

Gaussian distribution, unknown mean and precision (2/2)

- The conjugate prior distribution is the normal-gamma distribution

$$P(\mu, \lambda | \mu_0, \beta, a, b) = \text{Normal-Gamma}(\mu, \lambda | \mu_0, \beta, a, b)$$

- Posterior

$$P(\mu, \lambda | x) = \text{Normal-Gamma}(\mu, \lambda | \mu_n, \beta_n, a_n, b_n),$$

with

$$\mu_n = \frac{\beta \mu_0 + n\overline{x}}{\beta + n}$$

$$\beta_n = \beta + n$$

$$a_n = a + \frac{n}{2}$$

$$b_n = b + \frac{1}{2}\left(ns + \frac{\beta n (\overline{x} - \mu_0)^2}{\beta + n} \right)$$

Consistency

- If $p(x|\theta_t)$ is the true data generating mechanism, and A is a neighborhood of θ_t, then

$$P(\theta \in A|x) \overset{n\to\infty}{\to} 1.$$

- The posterior distribution concentrates around the true value (if such a value exists!).
- It follows that

$$\overline{\theta}_{MAP} \overset{n\to\infty}{\to} \theta_t \quad \text{and} \quad \overline{\theta}_{ML} \overset{n\to\infty}{\to} \theta_t$$

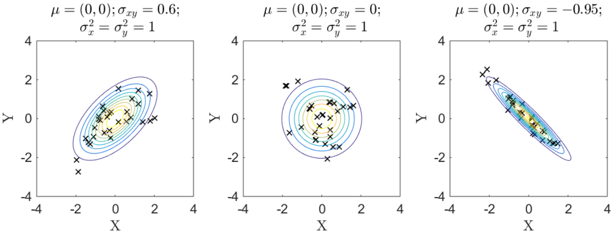
Multivariate Gaussian distribution

$$N_D(x|\mu,\Sigma) \equiv (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

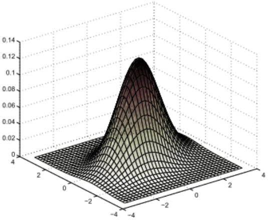
- D: dimension, μ: mean, Σ: covariance matrix. With $D = 2$:

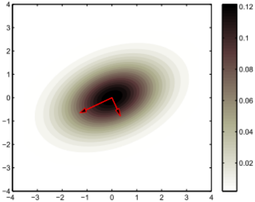
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- $\sigma_{12} = \sigma_{21}$: covariance between x_1 and x_2. (tells direction of dependency)
- $\rho_{12} = \sigma_{12}/(\sigma_1\sigma_2)$:correlation between x_1 and x_2. (direction and strength)



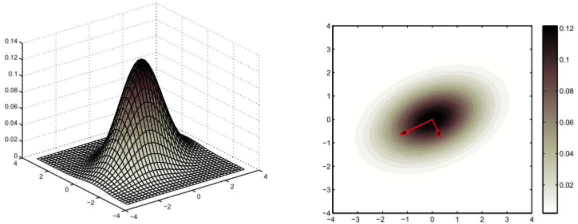
Multivariate Gaussian - characterization (1/2)*





- Eigendecomposition
$$\Sigma = E\Lambda E^{-1},$$
where $E^T E = I$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_D)$.
- Now the transformation
$$y = \Lambda^{-\frac{1}{2}} E^T (x - \mu)$$
can be shown to have the distribution $N_D(0, I)$ (product of D independent standard Gaussians)

Multivariate Gaussian - characterization (2/2)*



- Thus, $x = E\Lambda^{\frac{1}{2}} y + \mu$ with distribution $N_D(\mu, \Sigma)$ is obtained from standard independent Gaussians y by
  - *scaling* by the square roots of eigenvalues
  - *rotating* by the eigenvectors
  - *shifting* by adding the mean

Marginalization and conditioning (1/2)

- Let $z \sim N(\mu, \Sigma)$ and consider partitioning it as:

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

with

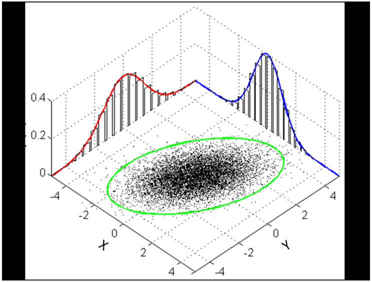
$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

Marginalization and conditioning (2/2)

- Then

$$P(x) \sim N(\mu_x, \Sigma_{xx}) \quad \textbf{(marginalization)}$$
$$P(x|y) = N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$
$$\textbf{(conditioning)}$$

$\Longrightarrow$Marginals and conditionals of M-V Gaussians are still M-V Gaussian.



Properties of multivariate Gaussian

- **Linear transformation**: if

$$y = Mx + \eta,$$

where $x \sim N(\mu_x, \Sigma_x)$ and $\eta \sim N(\mu, \Sigma)$,then

$$P(y) = N(y|M\mu_x + \mu, M\Sigma_x M^T + \Sigma)$$

- **Completing the square**:

$$\frac{1}{2}x^T Ax - b^T x = \frac{1}{2}(x - A^{-1}b)^T A(x - A^{-1}b) - \frac{1}{2}b^T A^{-1} b$$

From which one can derive, for example

$$\int \exp(-\frac{1}{2}x^T Ax + b^T x)dx = \sqrt{\det(2\pi A^{-1})}\exp(\frac{1}{2}b^T A^{-1} b)$$

Gaussian posterior

- General recipe:
  - Start with the expression of log-posterior $\log P(\theta \,|\, D)$
  - Drop all terms that do not depend on the parameters θ
  - Recognize that the expression is a quadratic function of the parameters:
    $$\frac{1}{2}\theta^T A\theta + b^T\theta$$
  - Use "completing the square" to find the parameters $\mu = A^{-1}b$ and $\Sigma = A^{-1}$
  - Conclude that the posterior follows Gaussian $P(\theta \,|\, D) = N(\theta \,|\, \mu, \Sigma)$

Multivariate Gaussian - Bayesian learning

- Gaussian-Wishart is the conjugate prior, when $X_i \sim N(\mu, \Lambda)$ and both mean μ and precision Λ are unknown:

$$P(\mu, \Lambda | \mu_0, \beta, W, \nu) = N(\mu | \mu_0, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda | W, \nu)$$

- If X_i are scalar, this is equivalent to the Gaussian-Gamma distribution.
- Posterior

$$P(\mu, \Lambda | x) = N(\mu | \mu_n, (\beta_n \Lambda)^{-1}) \mathcal{W}(\Lambda | W_n, \nu_n)$$

Wishart distribution*

- Wishart distribution is a distribution for nonnegative-definite matrix-valued random variables

$$\Lambda \sim \mathcal{W}(\Lambda | W, \nu)$$

$$E(\Lambda) = \nu W$$
$$\mathrm{Var}(\Lambda_{ij}) = n(w_{ij}^2 + w_{ii} w_{jj})$$

Multivariate Gaussian distribution as a Bayesian network

- Every multivariate Gaussian distribution can be represented as a Bayesian network where each node is associated a univariate Gaussian whose mean is a linear combination of the of the values of the parents
- Edges in the skeleton = non-zero values in the precision matrix

What's next?

- More complex models
  - Linear models
  - Clustering
  - ...
- Cannot compute $P(\theta \mid D)$ in closed form $\Rightarrow$ need to use approximations
  - Laplace approximation
  - EM algorithm
  - Markov chain Monte Carlo
  - Variational inference

Further readings

- Bishop: 2.1-2.4