Statistics for Machine Learning

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

Representing Data

- Categorical (or nominal): the observed value belongs to one of a number of classes, with no intrinsic ordering. Example: There are 4 kinds of jobs: `soldier`, `sailor`, `tinker`, `spy`.
- Ordinal: the observed value belongs to one of a number of classes with an ordering or ranking of the classes. Example: `cold`, `cool`, `warm`, `hot`.
- Numerical: the observed value takes on values that are numerical. Example: the `temperature` measured by a thermometer; the `salary` that someone earns.

Probability Density Functions (PDF)

For a continuous variable x, the probability density function $p(x)$ is defined such that

$$p(x) \geq 0, \int_{-\infty}^{+\infty} p(x)dx = 1, p(a < x < b) = \int_{a}^{b} p(x)dx.$$

The PDFs are also referred as distributions.

Averages/Expectations

$$\langle f(x) \rangle_{p(x)}$$

denotes the average or expectation of $f(x)$ with respect to the distribution $p(x)$. When the context is clear, one may drop the notational dependency on $p(x)$: $\langle f(x) \rangle$. An alternative notation is

$$\mathbb{E}(x).$$

The notation

$$\langle f(x)|y \rangle$$

is shorthand for the average of $f(x)$ conditioned on knowing the state of variable y, i.e. the average of $f(x)$ with respect to the distribution $p(x|y)$.

Averages/Expectations

In the discrete case

$$\langle f(x) \rangle = \sum_{\mathsf{x}} f(x = \mathsf{x}) p(x = \mathsf{x})$$

and in the continuous case

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x) p(x) dx.$$

Cumulative Distribution Function (CDF)

For a univariate distribution $p(x)$, the cumulative distribution function is defined as

$$cdf(y) = p(x \leq y) = \langle \mathbb{I}[x \leq y] \rangle_{p(x)}$$

with $cdf(-\infty) = 0$ and $cdf(+\infty) = 1$.

Moments

The kth moment of a distribution is given by the average of x^k under the distribution:

$$\langle x^k \rangle_{p(x)}.$$

For $k = 1$, we have the mean, typically denoted by μ,

$$\mu \equiv \langle x \rangle.$$

Moment Generating Function

For a distribution $p(x)$, we define the moment generating function $g(t)$ as

$$g(t) = \langle e^{tx} \rangle_{p(x)}.$$

The usefulness of this is that by differentiating $g(t)$, it is possible to generate the moments

$$\lim_{t \to 0} \frac{d^k}{dt^k} g(t) = \langle x^k \rangle_{p(x)}.$$

Mode

The mode x_* of a distribution $p(x)$ is the state of x at which the distribution takes its highest value,

$$x_* = \arg\max p(x).$$

A distribution could have more than one mode (be multi-modal).

Variance

The variance is given by

$$\sigma^2 \equiv \langle (x - \langle x \rangle)^2 \rangle_{p(x)}.$$

The square root of the variance, σ, is called the standard deviation. The notation $\mathrm{var}(x)$ is also used to emphasise for which variable the variance is computed. An equivalent expression is

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2.$$

For a multivariate distribution the matrix with elements

$$\Sigma_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle,$$

where $\mu_i = \langle x_i \rangle$, is called the covariance matrix. The diagonal entries of the covariance matrix contain the variance of each variable. An equivalent expression is

$$\Sigma_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle.$$

Correlation

The correlation matrix has elements

$$\rho_{ij} = \left\langle \frac{x_i - \mu_i}{\sigma_i} \frac{x_j - \mu_j}{\sigma_j} \right\rangle,$$

where σ_i is the standard deviation of variable x_i. The correlation is a normalized form of the covariance so that each element is bounded $-1 \leq \rho_{ij} \leq 1$.

Skewness and Kurtosis

The skewness is a measure of the asymmetry of a distribution:

$$\gamma_1 \equiv \frac{\langle (x - \langle x \rangle)^3 \rangle_{p(x)}}{\sigma^3}.$$

A positive skewness means the distribution has a heavy tail to the right. Similarly, a negative skewness means the distribution has a heavy tail to the left.
The kurtosis is a measure of how peaked around the mean a distribution is:

$$\gamma_2 \equiv \frac{\langle (x - \langle x \rangle)^4 \rangle_{p(x)}}{\sigma^4} - 3.$$

A distribution with positive kurtosis has more mass around its mean than would a Gaussian with the same mean and variance. These are also called super Gaussian. Similarly a negative kurtosis (sub Gaussian) distribution has less mass around its mean than the corresponding Gaussian. The kurtosis is defined such that a Gaussian has zero kurtosis (which accounts for the -3 term in the definition).

Dirac delta

For continuous x, we define the Dirac Delta function

$$\delta(x - x_0)$$

which is zero everywhere except at x_0, where there is a spike. In addition,

$$\int_{-\infty}^{+\infty} \delta(x - x_0)dx = 1$$

and

$$\int_{-\infty}^{+\infty} \delta(x - x_0)f(x)dx = f(x_0).$$

One can view the Dirac delta function as an infinitely narrow Gaussian:

$$\delta(x - x_0) = \lim_{\sigma \to 0} \mathcal{N}(x|x_0, \sigma^2).$$

Kronecker delta

The Kronecker Delta function

$$\delta_{x,x_0}$$

is zero everywhere except at x_0, where $\delta_{x_0,x_0} = 1$. The Kronecker delta is equivalent to $\delta_{x,x_0} = \mathbb{I}[x = x_0]$.

Attention: notice that the expression $\delta(x, x0)$ is used to denote either the Dirac or Kronecker delta, depending on the context.

Empirical Distribution

Consider a sample set of datapoints $x^1, ..., x^N$, which are states of a random variable x.

For a discrete variable x the empirical distribution is

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \delta_{x,x^n}.$$

For a continuous variable x the empirical distribution is

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x^n).$$

Empirical Distribution

The mean of the empirical distribution is given by

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x^n.$$

The variance of the empirical distribution is given by

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x^n - \hat{\mu})^2.$$

Empirical Distribution

For vectors, the mean of the empirical distribution is given by

$$\hat{\mu}_i = \frac{1}{N} \sum_{n=1}^{N} x_i^n.$$

For vectors, the variance of the empirical distribution is given by

$$\hat{\Sigma}_{ij} = \frac{1}{N-1} \sum_{n=1}^{N} (x_i^n - \hat{\mu}_i)(x_j^n - \hat{\mu}_j).$$

The Kullback-Leibler Divergence

For two distributions $q(x)$ and $p(x)$, the Kullback-Leibler divergence is given by

$$\mathrm{KL}(q|p) \equiv \langle \log q(x) - \log p(x) \rangle_{q(x)}.$$

The Kullback-Leibler divergence measures the 'difference' between distributions q and p.

$\mathrm{KL}(q|p) \geq 0$

Consider the following linear bound on the function $\log(y)$

$$\log(y) \leq y - 1.$$

Then,

$$\frac{p(x)}{q(x)} - 1 \geq \log \frac{p(x)}{q(x)}.$$

Since probabilities are non-negative, we can multiply both sides by $q(x)$ to obtain

$$p(x) - q(x) \geq q(x) \log p(x) - q(x) \log q(x).$$

We now integrate (or sum in the case of discrete variables) both sides

$$1 - 1 \geq \langle \log p(x) - \log q(x) \rangle_{q(x)}.$$

Rearranging gives

$$\mathrm{KL}(q|p) \equiv \langle \log q(x) - \log p(x) \rangle_{q(x)} \geq 0.$$

Entropy

The entropy is defined as

$$H(p) \equiv -\langle \log p(x) \rangle_{p(x)}.$$

The entropy is a measure of the uncertainty in a distribution. The more similar p is to a uniform distribution, the greater will be the entropy.

Mutual Information

The mutual information is a measure of dependence between (sets of) variables $\mathcal{X}$ and $\mathcal{Y}$, conditioned on variables $\mathcal{Z}$:

$$\mathsf{MI}(\mathcal{X};\mathcal{Y}|\mathcal{Z}) \equiv \langle \mathsf{KL}(p(\mathcal{X},\mathcal{Y}|\mathcal{Z})|p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}))\rangle_{p(\mathcal{Z})} \geq 0.$$

If $\mathcal{X} \perp\!\!\!\perp Y|\mathcal{Z}$ is true, then $\mathsf{MI}(\mathcal{X};\mathcal{Y}|\mathcal{Z})$ is zero, and vice-versa. When $\mathcal{Z} = \varnothing$, the average over $p(\mathcal{Z})$ is absent and one writes $\mathsf{MI}(\mathcal{X};\mathcal{Y})$.

Some Distributions

Bernoulli Distribution

The Bernoulli distribution concerns a discrete binary variable x, with $\mathrm{dom}(x) = \{0, 1\}$.

$$p(x = 1) = \theta, p(x = 0) = 1 - \theta, \theta \in [0, 1].$$

In addition,

$$\langle x \rangle = \theta$$

and

$$\mathsf{var}(x) = \theta(1 - \theta).$$

Some Distributions

Categorical Distribution

The categorical distribution generalizes the Bernoulli distribution to more than two (symbolic) states. For a discrete variable x, with symbolic states $\mathrm{dom}(x) = \{1, \ldots, C\}$,

$$p(x = i) = \theta_i, \sum_{i=1}^{C} \theta_i = 1, \theta_i \in [0, 1].$$

The Dirichlet distribution is conjugate to the categorical distribution.

Some Distributions

Binomial Distribution

The probability that in n Bernoulli Trials (independent samples), k 'success' states 1 will be observed is

$$p(y = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k},$$

where

$$\binom{n}{k} \equiv \frac{n!}{k!(n-k)!}.$$

In addition,

$$\langle y \rangle = n\theta$$

$$\text{var}(y) = n\theta(1-\theta).$$

The Beta distribution is the conjugate prior for the Binomial distribution.

Some Distributions

Multinomial Distribution

In n Categorical Trials (independent samples), the probability of observing the state 1 y_1 times, state 2 y_2 times, ..., state C y_C times is

$$p(y|\theta) = \frac{n!}{y_1! \dots y_C} \prod_{i=1}^{C} y_i,$$

where

$$\sum_{i=1}^{C} y_i = n.$$

In addition,

$$\langle y_i \rangle = n\theta_i, \mathsf{var}(y_i) = n\theta_i(1-\theta_i), \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle = -n\theta_i\theta_j \ (i \neq j).$$

The Dirichlet distribution is the conjugate prior for the multinomial distribution.

Some Distributions

Poisson Distribution

The Poisson distribution can be used to model situations in which the expected number of events scales with the length of the interval within which the events can occur. If λ is the expected number of events per unit interval, then the distribution of the number of events x within an interval $t\lambda$ is

$$p(x = k|\lambda) = \frac{1}{k!}e^{-\lambda t}(\lambda t)^k, k = 0, 1, 2, \ldots.$$

For a unit length interval ($t = 1$),

$$\langle x \rangle = \lambda, \mathsf{var}(x) = \lambda.$$

Some Distributions

Uniform Distribution

For a variable x, the distribution is uniform if $p(x)$ is constant over the domain of the variable.

Exponential Distribution

For $x \geq 0$,

$$p(x|\lambda) = \lambda e^{-\lambda x},$$

where $\lambda \geq 0$. In addition,

$$\langle x \rangle = \frac{1}{\lambda}, \mathrm{var}(x) = \frac{1}{\lambda^2}$$

and $b = 1/\lambda$ is called the scale.

Some Distributions

Gamma Distribution

$$\text{Gam}(x|\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}\begin{pmatrix} x \\ \beta \end{pmatrix}^{\alpha-1} e^{-\frac{x}{\beta}}, x \geq 0, \alpha > 0, \beta > 0,$$

α is called the shape parameter, β is the scale parameter and the Gamma function is defined as

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt.$$

The parameters are related to the mean, μ, and variance, σ^2, through

$$\alpha = \left(\frac{\mu}{\sigma}\right)^2, \beta = \frac{\sigma^2}{\mu}.$$

Some Distributions

Inverse Gamma Distribution

$$\text{InvGam}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{1}{x^{\alpha+1}}e^{-\beta/x}.$$

In addition,

$$\langle x\rangle = \frac{\beta}{\alpha-1}, \alpha > 1$$

and

$$\text{var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2.$$

Some Distributions

Beta Distribution

$$p(x|\alpha,\beta) = \frac{1}{\mathsf{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \leq x \leq 1.$$

where the Beta function is defined as

$$\mathsf{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

In addition,

$$\langle x \rangle = \frac{\alpha}{\alpha+\beta}, \mathsf{var}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Some Distributions

Laplace Distribution

$$p(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}.$$

In addition,

$$\langle x \rangle = \mu, \mathsf{var}(x) = 2b^2.$$

The Laplace distribution is also known as the Double Exponential distribution.

Some Distributions

Univariate Gaussian or Normal Distribution

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

In addition,

$$\langle x \rangle = \mu, \mathsf{var}(x) = \sigma^2.$$

For $\mu = 0$ and $\sigma = 1$, the Gaussian is called the standard normal distribution.

Some Distributions

Student's t-distribution

$$p(x|\mu,\lambda,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\nu\pi}\right)^{\frac{1}{2}}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}},$$

where ν represents the degrees of freedom and λ scales the distribution. In addition,

$$\langle x\rangle = \mu, \text{var}(x) = \frac{\nu}{\lambda(\nu-2)} \text{ with } \nu > 2.$$

For $\nu \to \infty$ the distribution tends to a Gaussian with mean μ and variance $1/\lambda$. As ν decreases the tails of the distribution become fatter.

Some Distributions

Dirichlet Distribution

The Dirichlet distribution is a distribution on probability distributions:

$$p(\mathbf{x}|\alpha) = \mathsf{Dirichlet}(\mathbf{x}|\alpha) = \frac{1}{\mathsf{Z}(\alpha)} \prod_{i=1}^{Q} x_i^{\alpha_i - 1},$$

where $Q \geq 2$, $\sum_{i=1}^{Q} x_i = 1$ and $\alpha_i > 0, x_i \geq 0, \forall i \in \{1, \ldots, Q\}$. $\{x_i\}_{i=1}^{Q}$ is the standard $Q-1$ simplex and $\mathsf{Z}(\alpha) = \prod_{i=1}^{Q} \Gamma(\alpha_i)/\Gamma(\alpha_0)$. In addition,

$$\langle x_i \rangle = \frac{\alpha_i}{\alpha_0}, \mathsf{var}(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)},$$

$$\mathsf{cov}(x_i, x_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}, i \neq j, \alpha_0 = \sum_{i=1}^{Q} \alpha_i.$$

In the binary case $Q = 2$, this is equivalent to a Beta distribution.

Some Distributions

Multivariate Gaussian or Normal Distribution

$$p(\mathbf{x}|\mu,\mathbf{\Sigma}) = \mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)},$$

where μ is the mean vector of the distribution and $\mathbf{\Sigma}$ is the covariance matrix. In addition,

$$\langle \mathbf{x} \rangle_{\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma})} = \mu, \left\langle (\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathsf{T}} \right\rangle_{\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma})} = \mathbf{\Sigma}.$$

Exponential Family

The exponential family contains many standard distributions, including the Gaussian, Gamma, Poisson, Dirichlet, Multinomial, ... For a distribution on a (possibly multidimensional) variable x (continuous or discrete) an exponential family model is of the form

$$p(x|\theta) = h(x)\exp(\sum_i \eta_i(\theta)T_i(x) - \psi(\theta)),$$

where θ are the parameters, T_i is a function and ψ is a function that ensure normalization

$$\psi(\theta) = \log \int_x h(x)\exp(\sum_i \eta_i(\theta)T_i(x)).$$

Canonical Form

Setting $\eta(\theta) = \theta$, the canonical exponential family model is given by

$$p(x|\theta) = h(x)\exp(\theta^{\mathsf{T}}\mathbf{T}(x) - \psi(\theta)).$$

Exponential Family

Example

The univariate Gaussian belongs to the exponential family

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = e^{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)}.$$

A possible definition is

$$t_1(x) = x, t_2(x) = -\frac{x^2}{2}, \theta_1 = \mu, \theta_2 = \sigma^2, h(x) = 1,$$

$$\eta_1(\theta) = \frac{\theta_1}{\theta_2}, \eta_2(\theta) = \frac{1}{\theta_2}, \psi(\theta) = \frac{1}{2}\left(\frac{\theta_1^2}{\theta_2} + \log(2\pi\theta_2)\right).$$

Exponential Family

Conjugate Priors

Consider the canonical exponential family likelihood

$$p(x|\theta) = h(x)e^{\theta^\top \mathbf{T}(x) - \psi(\theta)}.$$

The conjugate prior with hyperparameters α, γ is given by

$$p(\theta|\alpha, \gamma) \propto e^{\theta^\top \alpha - \gamma\psi(\theta)}$$

and the posterior is given by

$$p(\theta|x) \propto p(x|\theta)p(\theta) \propto e^{\theta^\top (\mathbf{T}(x)+\alpha) - (\gamma+1)\psi(\theta)}.$$

If the posterior distribution is in the same family as the prior distribution, the prior and posterior are called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Learning Distributions

For a distribution $p(x|\theta)$, parameterized by θ, and data $\mathcal{X} = \{x^1 \ldots, x^N\}$, learning corresponds to inferring the θ that 'best fits' the data $\mathcal{X}$.

- Bayesian Methods: makes use of the relation $p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)$. This gives rise to a distribution of the parameters θ.
- Maximum a Posteriori (MAP): summarizes the posterior using

$$\theta^{\mathsf{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{X}).$$

- Maximum Likelihood: maximizes the likelihood of observing the data:

$$\theta^{\mathsf{ML}} = \arg\max_{\theta} p(\mathcal{X}|\theta).$$

- Moment Matching: based on an empirical estimate of a moment (say the mean), θ is set such that the moment (or moments) of the distribution matches the empirical moment.
- Pseudo Likelihood: provides an approximation to the likelihood function of a set of observed data which may either provide a computationally simpler problem for estimation, or may provide a way of obtaining explicit estimates of model parameters.

Learning a Gaussian

Maximum Likelihood Training

Given a set of training data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, drawn from a Gaussian $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$ with unknown mean μ and covariance $\mathbf{\Sigma}$, how can we find these parameters?
Assuming the data are drawn i.i.d., the loglikelihood is

$$L(\mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \log p(\mathbf{x}|\mu, \mathbf{\Sigma}) = -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}^n - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}^n - \mu) - \frac{N}{2}\log\det(2\pi\mathbf{\Sigma}).$$

Optimal μ

$$\nabla_\mu L(\mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \mathbf{\Sigma}^{-1}(\mathbf{x}^n - \mu)$$

$$\nabla_\mu L(\mu^*, \mathbf{\Sigma}) = \mathbf{0} \Leftrightarrow \sum_{n=1}^{N} \mathbf{\Sigma}^{-1}\mathbf{x}^n = N\mu^*\mathbf{\Sigma}^{-1} \Leftrightarrow \mu^* = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^n.$$

Learning a Gaussian

Optimal Σ It is possible to write the loglikelihood as follows

$$L(\mu, \mathbf{\Sigma}) = -\frac{1}{2}\text{trace}\left(\mathbf{\Sigma}^{-1}\sum_{n=1}^{N}(\mathbf{x}^n - \mu)(\mathbf{x}^n - \mu)^{\mathsf{T}}\right)$$

$$+\frac{N}{2}\log\det(\mathbf{\Sigma}^{-1}) - \frac{N}{2}\log[(2\pi)^m],$$

where m is the dimension of $\mathbf{x}$. Remember the 'Matrix Calculus' slide from 'Linear Algebra Introduction',

$$\det(\mathbf{\Sigma}) = \det^{-1}(\mathbf{\Sigma}^{-1})$$

and

$$\mathbf{M} \equiv \sum_{n=1}^{N}(\mathbf{x}^n - \mu)(\mathbf{x}^n - \mu)^{\mathsf{T}}, \mathbf{M} = \mathbf{M}^{\mathsf{T}}.$$

Learning a Gaussian

Then,

$$\frac{\partial}{\partial \mathbf{\Sigma}^{-1}} L(\mu, \mathbf{\Sigma}^{-1}) = -\frac{1}{2}\mathbf{M} + \frac{N}{2}\mathbf{\Sigma}.$$

Finally,

$$\frac{\partial}{\partial \mathbf{\Sigma}^{-1}} L(\mu, \mathbf{\Sigma}^{*-1}) = 0 \Leftrightarrow \mathbf{\Sigma}^{*} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^n - \mu)(\mathbf{x}^n - \mu)^{\mathsf{T}}.$$

In fact, these equations simply set the parameters to their sample statistics of the empirical distribution.

Learning a Gaussian

Bayesian inference of the mean and variance for the univariate case

Assuming i.i.d. data, $\mathcal{X}$, the likelihood is

$$p(\mathcal{X}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x^n-\mu)^2}.$$

The posterior of the parameters is

$$p(\mu,\sigma^2|\mathcal{X}) \propto p(\mathcal{X}|\mu,\sigma^2)p(\mu,\sigma^2).$$

For a Gauss-Inverse-Gamma prior

$$p(\mu,\sigma^2) = \mathcal{N}(\mu|\mu_0,\gamma\sigma^2)\mathsf{InvGam}(\sigma^2|\alpha,\beta)$$

the posterior is also Gauss-Inverse-Gamma

$$p(\mu,\sigma^2|\mathcal{X}) = \mathcal{N}\left(\mu\left|\frac{\tilde{b}}{\tilde{a}},\frac{\sigma^2}{\tilde{a}}\right.\right)\mathsf{InvGam}\left(\sigma^2\left|\alpha+\frac{N}{2},\beta+\frac{1}{2}\left(\tilde{c}-\frac{\tilde{b}^2}{\tilde{a}}\right)\right.\right),$$

Learning a Gaussian

with

$$\tilde{a} = \frac{1}{\gamma} + N, \tilde{b} = \frac{\mu_0}{\gamma} + \sum_{n=1}^{N} x^n, \tilde{c} = \frac{\mu_0^2}{\gamma} + \sum_{n=1}^{N} (x^n)^2.$$

Matlab

In the Gaussian distribution, it is common to use the precision parameter defined as the inverse variance:

$$\lambda \equiv \frac{1}{\sigma^2}.$$

Then, the prior is Gauss-Gamma

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, \gamma\lambda^{-1})\text{Gam}(\lambda|\alpha, \beta)$$

and the posterior is also Gauss-Gamma

$$p(\mu, \lambda|\mathcal{X}) = \mathcal{N}\left(\mu \left| \frac{\tilde{b}}{\tilde{a}}, \frac{1}{\tilde{a}\lambda} \right.\right) \text{Gam}\left(\lambda \left| \alpha + \frac{N}{2}, \left[\frac{1}{\beta} + \frac{1}{2}\left(\tilde{c} - \frac{\tilde{b}^2}{\tilde{a}}\right)\right]^{-1} \right.\right).$$

```
>>setup;
>>demoGaussBayes;
```