Notes 2-Probability Distribution

Junda Du

2019-8-12

1 Binary Variables

Bernoulli Distribution Points:

1. Definition: $x \in \{0,1\}, p(x=1|\mu) = \mu$, the probability distribution over x is $Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$,which is known as *Bernoulli* distribution. It's normalized and easy to get $E[x] = \mu, var[x] = \mu(1-\mu)$.
2. ML to estimate parameters given data set D(**default** :$\{x_1, \ldots, x_N\}$ iid):
   - $p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$
   - $ln\, p(D|\mu) = \sum_{n=1}^{N} ln\, p(x_n|\mu) = \sum_{n=1}^{N} \{x_n ln\, \mu + (1-x_n) ln(1-\mu)\}$
   - Set the derivative of $ln\, p(D|\mu)$ with respect to μ equal to zero, then obtain ML estimator $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$
3. Bayesian treatment for the over-fitting problem given small data sets.
   - Introduce a conjugate prior distribution $p(\mu) \propto \mu^a(1-\mu)^b$ according to $p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$[Same form for prior variable, const $a, b, c, \cdots$ for $f(x_n)$ or non-prior variables], so that the posterior probability $p(\mu|D)$ will have the same functional form as the prior probablity.
   - Choose a prior $p(\mu) = Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, which is normalized by $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. Its $E[\mu] = \frac{a}{a+b}, var[\mu] = ab(a+b)^2(a+b+1)$.
   - The posterior probability $p(\mu|m,l,a,b) \propto \mu^{m+a-1}\mu^{l+b-1}, l = N-m$, then normalizing it according to the form of the prior probability, get $p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(m+b)}\mu^{m+a-1}\mu^{l+b-1}, l = N-m$
   - When getting a new observation, we can take the old prior probability as the prior probability, and multiply it by the likelihood function for the new observation to get the current prior probablity. It's called the *sequential* approach and depends only on the assumption of i.i.d. data.

- Utilize the posterior probability to predict, $p(x=1|D) \overset{Baysian's theorem}{=\!=\!=} \int_0^1 p(x=1|\mu)p(\mu|D)d\mu = \int_0^1 \mu p(\mu|D)d\mu = E[\mu|D] \overset{Beta\,distribution's\,identity}{=\!=\!=} \frac{m+a}{m+a+l+b}$

4. Show that **generally**, as we observe more and more data, the uncertainty represented by posterior distribution **on average** will steadily decrease.
   First, let take the mean for example, the prior mean $E_\theta[\theta] = \int \theta p(\theta)d\theta$, the posterior mean on average $E_D[E_\theta[\theta|D]] = \int\{\int \theta p(\theta|D)d\theta\}p(D)dD$, easily get $E_\theta[\theta] = E_D[E_\theta[\theta|D]]$
   Similarly, we can show that $var_\theta[\theta] = E_D[var_\theta[\theta|D]] + var_D[E_\theta[\theta|D]]$,, and variance ≥ 0 .

Binomial Distribution Points:

1. Definition: Given the data set D $[p(x_i|\mu = Bern(x_i|\mu)]$,$m$ is the number of the observations of $x=1$.
   The probability distribution over m $p(m|N,\mu) \propto \mu^m(1-\mu)^{N-m}$.
   After normalization,

$$p(m|N,\mu) = Bin(m|N,\mu) = ?(\frac{N}{m})\mu^m(1-\mu)^{N-m}, (\frac{N}{m}) = \frac{N!}{(N-m)!m!}$$

2. $E[m] \overset{definition}{=\!=\!=} \sum_{m=0}^{N} mBin(m|N,\mu) \overset{m=x_1+\cdots+x_N}{=\!=\!=} \sum_{n=0}^{N} E[x_n] = N\mu$
   $var[m] \overset{definition}{=\!=\!=} E[(m-E[m])^2] = \sum_{m=0}^{N}(m-E[m])^2 Bin(m|N,\mu)$
   $\overset{m=x_1+\cdots+x_N,iid}{=\!=\!=\!=} \sum_{n=0}^{N} var[x_n] = N\mu(1-\mu)$

2 Multinomial Variables

1. Definition: K-dimensional vector $\boldsymbol{x}$ only have one element $x_k = 0$, all remaining elements $= 0$, $p(x_k = 1) = \mu_k, \sum_k \mu_k = 1$. The probability over $\boldsymbol{x}$ is $p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \sum_k \mu_k = 1$. Easily get $\sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1, E[\boldsymbol{x}|\boldsymbol{\mu}] = \sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\mu})\boldsymbol{x} = (\mu_1,\ldots,\mu_M)^T = \boldsymbol{\mu}$

2. ML to estimate parameters

   - $p(D|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_n k} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}, m_k = \sum_n x_{nk}$
   - $ln\, p(D|\boldsymbol{\mu}) + \lambda(\sum_k \mu_k - 1) = \sum_{k=1}^{K} m_k ln\, \mu_k + \lambda(\sum_k \mu_k - 1)$
   - Set the derivative with respect to μ_k to zero, obtain $\mu_k = -\frac{m_k}{\lambda}$, substituting it into constraint $\sum_k \mu_k = 1$ to give $\lambda = -N$, so $\mu_k^{ML} = \frac{m_k}{N}$

3. Bayesian's treatment

   - Conjugate prior $p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, 0 \leq \mu_k \leq 1, \sum_k \mu_k = 1$, according to $p(D|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{m_k}$
   - After normalization, $p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \alpha_0 = \sum_{k=1}^{K} \alpha_k$
   - $p(\boldsymbol{\mu}|D, \boldsymbol{\alpha}) \propto p(D|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$, then normalizing it according to the form of the prior probability, get $p(\boldsymbol{\mu}|D, \boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha} + \boldsymbol{m}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$

4. *Multinomial* distribution
   $Mult(m_1, \ldots, m_K|\boldsymbol{\mu}, N) \propto p(D|\boldsymbol{\mu}) \xrightarrow{Normalizing} (\frac{N}{m_1 m_2 \ldots m_K}) \prod_{k=1}^{K} \mu_k^{m_k}, \sum_k m_k = N, (\frac{N}{m_1 m_2 \ldots m_K}) = \frac{N!}{m_1! \ldots m_K!}$

3 The Gaussion Distribution

Points:

- Definition of Gaussion distribution:
  $N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\}$ for univariable,
  $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\}$ for multivaribale.

- Situations it arises:
  A single variable's distribution that maximizes the entropy is the Gaussian.
  The sum of a set of random variables has a Gaussian distribution.

- Manipulating Gaussian distributions:

  – Represent Gaussian in a simpler form:

    1. We focus on the part dependent on $\boldsymbol{x}$: $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$, which is a quadratic form. Δ is called *Mahalanobis distance* from $\boldsymbol{\mu}$ to $\boldsymbol{x}$, and reduces to Euclidean distance when $\boldsymbol{\Sigma}$ is the identity matrix.
    2. Consider eigenvector equation for $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} \boldsymbol{u}_i = \lambda_i u_i$, where $i = 1, \cdots, D$ and eigenvector can be chosen to form an orthonormal set because $\boldsymbol{\Sigma}$ must be real and symmetric?. So $\boldsymbol{u}_i^T \boldsymbol{u}_j = I_{ij}$.

3. Then $\mathbf{\Sigma}, \mathbf{\Sigma}^{-1}$ can be expressed as these form: $\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T, \mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T$.
4. So the quadratic form becomes $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}, y_i = \boldsymbol{u}_i^T(\boldsymbol{x} - \boldsymbol{\mu})$.
5. In going from the $\boldsymbol{x}$ to the $\boldsymbol{y}$ coordinate system, we have a Jacobian matrix $\boldsymbol{J}$ with elements given by $J_{ij} = \frac{\delta x_i}{\delta y_j} = U_j i$, because $\boldsymbol{y} = \boldsymbol{U}(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{U} = (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_D^T)^T$, thus $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$. So $|\boldsymbol{J}|^2 = |\boldsymbol{U}^T|^2 = |\boldsymbol{U}^T||\boldsymbol{U}| = |\boldsymbol{U}^T\boldsymbol{U}| = |\boldsymbol{I}| = 1$. Also $|\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}$.
6. Finally $p(\boldsymbol{y}) = p(\boldsymbol{x})|\boldsymbol{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} exp\{-\frac{y_j^2}{2\lambda_j}\}$, which is the product of D independent univariate Gaussian distributions.
7. Provide an interpretation of the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$:

– Let's look at the moments and the covariance of the Gaussian distribution.

1. For first order moments/expectation:

$$\begin{aligned} E[\boldsymbol{x}] =& \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}\boldsymbol{x} d\boldsymbol{x} \\ \overset{\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu}}{=\!=\!=} & \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int exp\{-\frac{1}{2}\boldsymbol{z}^T \Sigma^{-1}\boldsymbol{z}\}(\boldsymbol{z} + \boldsymbol{\mu}) d\boldsymbol{z} \\ \overset{even,odd,symmetry}{=\!=\!=\!=\!=\!=\!=\!=\!=} & \boldsymbol{\mu} \end{aligned}$$

2. For second order moments:

$$\begin{aligned} E[\boldsymbol{x}\boldsymbol{x}^T] =& \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}\boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} \\ \overset{\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu}}{=\!=\!=} & \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int exp\{-\frac{1}{2}\boldsymbol{z}^T \Sigma^{-1}\boldsymbol{z}\}(\boldsymbol{z} + \boldsymbol{\mu})(\boldsymbol{z} + \boldsymbol{\mu})^T d\boldsymbol{z} \\ \overset{even,odd,symmetry}{=\!=\!=\!=\!=\!=\!=\!=\!=} & \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int exp\{-\frac{1}{2}\boldsymbol{z}^T \Sigma^{-1}\boldsymbol{z}\}(\boldsymbol{z}\boldsymbol{z}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T) d\boldsymbol{z} \\ \overset{\substack{\boldsymbol{z}=\boldsymbol{U}^{-1}\boldsymbol{y}=\boldsymbol{U}^T\boldsymbol{y}=\sum_{j=1}^{D} y_j\boldsymbol{u}_j \\ \mathbf{\Sigma}^{-1}=\sum_{i=1}^{D}\frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T}}{=\!=\!=\!=\!=\!=\!=\!=\!=} & \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \sum_{i=1}^{D}\sum_{j=1}^{D} \boldsymbol{u}_i\boldsymbol{u}_j^T \int exp\{-\sum_{k=1}^{D}\frac{y_k^2}{2\lambda_k}\} y_i y_j \\ \overset{\substack{\boldsymbol{u}_i^T\boldsymbol{u}_j=I_{ij}, |\mathbf{\Sigma}|^{1/2}=\prod_{j=1}^{D}\lambda_j^{1/2} \\ split\ i\ from\ the\ integral}}{=\!=\!=\!=\!=\!=\!=\!=\!=} & \boldsymbol{\mu}\boldsymbol{\mu}^T + \sum_{i=1}^{D} \boldsymbol{u}_i\boldsymbol{u}_i^T\lambda_i = \boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{\Sigma} \end{aligned}$$

3. For covariance:
   $cov[\boldsymbol{x}] = E[(\boldsymbol{x} - E[\boldsymbol{x}])(\boldsymbol{x} - E[\boldsymbol{x}])^T] = E[\boldsymbol{x}\boldsymbol{x}^T] - E[\boldsymbol{x}]E[\boldsymbol{x}]^T = \boldsymbol{\Sigma}$

4. As we can see:
   $\boldsymbol{\mu} = E[\boldsymbol{x}], \boldsymbol{\Sigma} = cov[\boldsymbol{x}] = E[(\boldsymbol{x} - E[\boldsymbol{x}])(\boldsymbol{x} - E[\boldsymbol{x}])^T]$

3.1 Conditional Gaussian distribution

We want to get $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$, and $p(\boldsymbol{x}) = p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ is Gaussian.We set

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

where $\boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^T$ because $\boldsymbol{\Sigma}$ is symmetric,$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Lambda}_{ab} = \boldsymbol{\Lambda}_{ba}^T$ because the inverse of a symmetric matrix is also symmetric, but $\boldsymbol{\Lambda}_{aa}$ is not simply given by $\boldsymbol{\Sigma}_{aa}^{-1}$.Next we will discuss the relationship between $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Sigma}_{aa}^{-1}$.
There is a identity for the inverse of a partitioned matrix,

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}, \boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}$$

The quantity $\boldsymbol{M}^{-1}$ is known as the *Schur complement* of the left-hand side with respect to $\boldsymbol{D}$.
Using the definition

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix},$$

we have $\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}, \boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}$

$p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is simply evaluated from the joint distribution $p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ by fixing $\boldsymbol{x}_b$ and then normalizing to obtain a valid probability over $\boldsymbol{x}_a$. We only consider the exponent part because normalization process is adapt coefficient to the exponent.

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = &-\frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &- \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ \xlongequal{fix\,\boldsymbol{x}_b, complete\,the\,square} &-\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} + const \end{aligned}$$

There is a technique called "**completing the square**", in which we obtain the required mean and variance by comparing given exponent with a general Gaussian distribution $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$'s exponent $-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const$

So we can obtain
$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$,
$\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$
we can also substitute $\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}, \boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}$
to obtain a new representation.

3.2 Marginal Gaussian distribution

Usually, we use $p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b$ to get marginal probablity over $\boldsymbol{x}_a$, so first we consider the terms involving $\boldsymbol{x}_b$ and then completing the square to facilitate integration.

$$\begin{aligned}
-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) =& -\frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\
& -\frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\
\xrightarrow{fix\ \boldsymbol{x}_a, complete\ the\ square} & -\frac{1}{2}\boldsymbol{x}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b + \boldsymbol{x}_b^T\boldsymbol{m} + \boldsymbol{R}(\boldsymbol{x}_a), \boldsymbol{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) \\
\xrightarrow{facilitate\ integration} & -\frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}) + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} + \boldsymbol{R}(\boldsymbol{x}_a)
\end{aligned}$$

Since $\frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}+\boldsymbol{R}(\boldsymbol{x}_a)$ is independent on $\boldsymbol{x}_b$, and $\int exp\{-\frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})d\boldsymbol{x}_b\}$ becomes the coefficient beacuse it's a normalized Gaussion.
So exponent part becomes

$$\begin{aligned}
\frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m} + \boldsymbol{R}(\boldsymbol{x}_a) =& \frac{1}{2}[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)]^T\boldsymbol{\Lambda}_{bb}^{-1}[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)] \\
& -\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b) + const \\
\xrightarrow{fix\ \boldsymbol{x}_b, complete\ the\ square} & -\frac{1}{2}\boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{x}_a + \\
& \boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a + const
\end{aligned}$$

So we can obtain
$\boldsymbol{\Sigma}_a = (\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})^{-1}$,
$\boldsymbol{\mu}_a = \boldsymbol{\Sigma}_a(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a = \boldsymbol{\mu}_a$
we can also substitute $\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa}-\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}, \boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa}-\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}$
to obtain $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{aa}$.

3.3 Bayes' theorem for Gaussian distribution

We want to obtain $p(\boldsymbol{y}), p(\boldsymbol{x}|\boldsymbol{y})$ given $p(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), p(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}, \boldsymbol{L}^{-1})$.

1. First, we can obtain $p(\boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{y})$ by $ln\, p(\boldsymbol{z}) = ln\, p(\boldsymbol{x}) + ln\, p(\boldsymbol{y}|\boldsymbol{x})$.
2. Second, writing $ln\, p(\boldsymbol{z})$ as the general Gaussion form $-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+const$ by considering first and second order item respectively.
3. Third, use conclusion of last two sections to obtain $p(\boldsymbol{y}), p(\boldsymbol{x}|\boldsymbol{y})$.

Ignoring the details, we can obtain
$p(\boldsymbol{y}) = N(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T)$,
$p(\boldsymbol{x}|\boldsymbol{y}) = N(\boldsymbol{x}|\boldsymbol{\Sigma}\{\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$

3.4 Maximum likelihood for the Gaussian

Given data set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$, assuming $\boldsymbol{x}_n$ are drawn independently from a multivariate Gaussian distribution, we want to estimate the parameters of the Gaussian distribution by maximum likelihood.

1. Write the log likelihood function.
2. Set the derivative with respect to $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ to zero respectively.Obtain

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$$

3. Evalute the expectation of estimated parameters $E[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu}, E[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N}\boldsymbol{\Sigma}$, then correct the bias obtaining

$$\boldsymbol{\Sigma} = \frac{N}{N-1}\boldsymbol{\Sigma}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$$

3.5 Sequential estimation

Robbins-Monro algorithm:If a problem is to find θ sastifying $f(\theta) = 0$, and $f(\theta)$ can be written as the form $f(\theta) = E[z|\theta] = \int zp(z|\theta)dz$, the root $\theta^{(N)} = \theta^{(N-1)} + a_{N-1}z(\theta^{(N-1)})$.
Take the maximum likelihood solution θ_{ML} for example.
By definition, θ_{ML} is a stationary point of the log likelihood function,

$$\frac{\partial}{\partial\theta}\left\{\frac{1}{N}\sum_{n=1}^{N}ln\,p(\boldsymbol{x}_n|\theta)\right\}\Bigg|_{\theta_{ML}} = 0.$$

It can be written as following form

$$\frac{\partial}{\partial\theta}\left\{\frac{1}{N}\sum_{n=1}^{N}ln\,p(\boldsymbol{x}_n|\theta)\right\} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}ln\,p(\boldsymbol{x}_n|\theta) = E_x\left[\frac{\partial}{\partial\theta}ln\,p(\boldsymbol{x}|\theta)\right] = 0$$

$z(\theta) = \frac{\partial}{\partial\theta}ln\,p(\boldsymbol{x}|\theta)$.Thus,

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1}z(\theta^{(N-1)}) = \theta^{(N-1)} + a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}ln\,p(\boldsymbol{x}_N|\theta^{(N-1)})$$

Take the sequential estimation of mean of a Gaussian distribution as a more specific example,

$$z(\mu^{(N-1)}) = \frac{\partial}{\partial\mu^{(N-1)}}ln\,p(\boldsymbol{x}_N|\mu^{(N-1)}, \sigma^2) = \frac{1}{\sigma^2}(\boldsymbol{x}_N - \mu^{(N-1)})$$

So we can undate μ by

$$\begin{aligned}\mu^{(N)} =&\mu^{(N-1)} + a_{N-1}z(\mu^{(N-1)})\\ =&\mu^{(N-1)} + a_{N-1}\frac{1}{\sigma^2}(\boldsymbol{x}_N - \mu^{(N-1)})\\ \xlongequal{a_{N-1}=\sigma^2/N}&\mu^{(N-1)} + \frac{1}{N}(\boldsymbol{x}_N - \mu^{(N-1)})\end{aligned}$$

3.6 Bayesian inference for the Gaussian

Considering a single Gaussian random varibale x, given a set of N observations $\boldsymbol{X} = \{x_1, \ldots, x_N\}$, the likelihood funciton is

$$p(\boldsymbol{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} exp\Big\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big\}$$

- Choose a conjugate prior $p(\mu)$ assuming variance σ^2 is known
  Conjugate prior $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$[Ignoring the process $\propto$], posterior $p(\mu|\boldsymbol{X}) \propto p(\boldsymbol{X}|\mu)p(\mu)$.
  By comleting the square[i.e.considering first and second order item], we can obtain $p(\mu|\boldsymbol{X}) = N(\mu|\mu_N, \sigma_N^2)$,

  $$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}, \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

  From sequnetial view of the inference problem, we can update $p(\mu|\boldsymbol{X})$ by
  $p(\boldsymbol{\mu}|\boldsymbol{X}) \propto \Big[p(\boldsymbol{\mu})\prod_{n=1}^{N-1} p(\boldsymbol{x}_n|\boldsymbol{\mu})\Big]p(\boldsymbol{x}_N|\boldsymbol{\mu})$

- Choose a conjugate prior $p(\lambda), \lambda = \frac{1}{\sigma^2}$ assuming mean μ is known Conjugate prior $p(\lambda) \propto \lambda^{a-1}exp(-b\lambda)$, after normalization $p(\lambda) = Gam(\lambda|a, b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}exp(-b\lambda)$.
  Then we can obtain $p(\lambda|\boldsymbol{X})$, which is also a gamma distribution.

- Choose a conjugate prior $p(\mu, \lambda)$
  The conjugate prior and posterior are both *normal-gamma* or *Gaussian-gamma* distribution.

3.7 Student's t-distribution

- We can get Student's t-distribution by integrating out the precision for univariate Gaussian multiplying the conjugate prior of precision(Gamma distribution).

- It alos can be obtained by adding up an infinite number of Gaussian distributions having the same mean but different precisions.

- It's more robust than Gaussion model,i.e. less sensitive to outliers.

3.8 Periodic variables

Here we'll talk about the *circular normal* or *von Mises* distribution. Let's how it's extended from two-variate Gaussian distribution.
Consider a Gaussian distribution over two varibales $\boldsymbol{x} = (x_1, x_2)$, having the mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and a covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, so

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} exp\Big\{-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2}\Big\}$$

Transforming form Cartesian coordinates to polar coordinates by $x_1 = rcos\theta, x_2 = rsin\theta$ also for $\boldsymbol{\mu}$ mapped by $\mu_1 = r_0 cos\theta_0, \mu_2 = r_0 sin\theta_0$
Then we obtain

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} exp\{m\, cos(\theta - \theta_0)\}, I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} exp\{m\, cos\theta\} d\theta.$$

The exponent is important, and $I_0(m)$ is the normalizaiton coefficient. We consider the maximum likelihood estimators for the parameters θ_0 and m for the Mises distribution.The likelihood function is given by $ln\, p(D|\theta_0, m) = -N\, ln(2\pi) - N\, ln\, I_0(m) + m \sum_{n=1}^{N} cos(\theta_n - \theta_0)$.

$$\frac{\partial ln\, p(D|\theta_0, m)}{\partial \theta_0} = 0 \Leftrightarrow \sum_{n=1}^{N} sin(\theta_n - \theta_0) = 0$$

$$\xLeftrightarrow{sin(A-B)=sin\, A\, cos\, B - cos\, A\, sin\, B} \theta_0^{ML} = tan^{-1}\Big\{\frac{\sum_n sin\, \theta_n}{\sum_n cos\, \theta_n}\Big\}$$

which is the result obtained for the mean of observations viewed in two-dimensional Cartesian space.?

$$\frac{\partial ln\, p(D|\theta_0, m)}{\partial m} = 0$$

$$\xLeftrightarrow{I_0'(m)=I_1'(m)} A(m) = \frac{1}{N}\sum_{n=1}^{N} cos(\theta_n - \theta_0^{ML}), A(m) = \frac{I_1(m)}{I_0(m)}$$

$$\Leftrightarrow A(m_{ML}) = \Big(\frac{1}{N}\sum_{n=1}^{N} cos\, \theta_n\Big) cos\, \theta_0^{ML} - \Big(\frac{1}{N}\sum_{n=1}^{N} sin\, \theta_n\Big) sin\, \theta_0^{ML}$$

3.9 Mixtures of Gaussians

4 The Exponent Family

A little hard...

5 Nonparametric Method

The nonparametric method focus mainly on frequentist method.

5.1 Histogram density

The probablity density of i_{th} box is obtained by $p_i = \frac{n_i}{N\Delta_i}$, where n_i is the number of oberservations of x falling in bin i, N is the total number of observations, and Δ_i is the width of i_{th} bin.

- Points:
  - When Δ is very small. the resulting density model is very spiky, with a lot of structure that is not present in the underlying distribution that generated the data set.Conversely, if Δ is too large, the result is too smooth and so fails to capture the bimodal property of the green curve. So Δ should be neither too small nor too large.
  - A histogram density model is also dependent on the choice of edge location for the bins.
- Advantages:
  - It's advantageous if the data set is large because once the histogram has been computed, the data set itself can be discarded.
  - It's easily applied if the data points are arriving sequentially.
  - It's useful for obtaining a quick visualization of data in one or two dimensions.
- Disadvantages:
  - The estimated density has discontinuities that are due to the bin edges ? rather than any property of the underlying distribution that generated the data.
  - Its scaling with dimensionality. In a space of high dimensionality D, the quantity of data needed to provide meaningful estimates of local probablity density and the calculation complexity would be prohibitive.(M_D)

5.2 Kernal density estimators

The probablity density is obtained by $p(\boldsymbol{x}) = \frac{K}{NV}$, where K is the number of oberservations falling in volume V around $\boldsymbol{x}$, N is the total number of observations.
The kernal approach fix V and determine K. There are some examples for kernal approach.

- Take the region R to be a small hypercube centred on the point $\boldsymbol{x}$.

$$p(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\Big(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\Big), k(\boldsymbol{u}) = \begin{cases} 1, & |u_i| \le \frac{1}{2}, i = 1, \dots, D \\ 0, & else \end{cases},$$

  where $h_D = V, K = \sum_{n=1}^{N} k\Big(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\Big)$, h is the smooth factor. It suffer from one of the same problems that the histogram method suffered from, namely the presence of artificial discontinuitied.

- A smoother density model by choosing a smoother kernal function Gaussian.

$$p(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}exp\Big\{\frac{\|\boldsymbol{x}-\boldsymbol{x}_n\|^2}{2h^2}\Big\}$$

- A general kernel function should satisfies:

$$k(\boldsymbol{u}) \ge 0, \int k(\boldsymbol{u})d\boldsymbol{u} = 1$$

5.3 Nearest-neighbor methods

The nearest-neighbor method fix K and determine V. It can be extended to classificaiton problem, which is specification of Baysian's estimation.