Machine Vision

November 9, 2018

1 Lecture 1

Bernoulli Distribution

$$Pr(x) = \lambda^x(1-\lambda)^{1-x}, \lambda \in= [0,1], x \in \{0,1\}$$

$$Pr(x) = Bern_x[\lambda]$$

Beta Distribution

$$Pr(\lambda) = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}, \alpha, \beta > 0$$
$$\Gamma(z) = \int_0^\infty t^{z-1}c^{-t}dt = (z-1)!$$
$$E[\lambda] = \frac{\alpha}{\alpha+\beta}$$
$$B(p,q) = \frac{q-1}{p+q+1}B(p,q-1)$$

α, β decide the coin fact λ

Categorical Distribution

$$Pr(x=k) = \lambda_k$$

$$Pr(x) = Cat_x[\lambda]$$

Dirichlet Distribution

$$Pr(\lambda_1 \dots \lambda_K) = \frac{\Gamma[\sum_{k=1}^K]\alpha_k}{\prod_{k=1}^K \Gamma[\alpha_k]}\prod_{k=1}^K \lambda_k^{\alpha_k - 1}$$
$$Pr(\lambda_1 \dots \lambda_K) = \mathrm{Dir}_{\lambda_1 \dots \lambda_K}[\alpha_1, \alpha_2 \dots, \alpha_K]$$

Univariate Normal Distribution

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-0.5(x-\mu)^2/\sigma^2]$$
$$Pr(x) = \mathrm{Norm}_x[\mu, \sigma^2]$$

Normal Inverse Gamma Distribution

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}\beta^{\alpha}}{\sigma\sqrt{2\pi}\Gamma[\alpha]}(\frac{1}{\sigma^2})^{\alpha+1} \exp[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}]$$
$$Pr(\mu, \sigma^2) = \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

Multivariate Normal Distribution

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\sum|^{1/2}} \exp[-0.5(\mathbf{x}-\mu)^T \sum^{-} 1(\mathbf{x}-\mu)]$$

Normal Inverse Wishart

$$Pr(\mu, \Sigma) = \frac{\gamma^{D/2}|\Psi|^{\alpha/2}|\Sigma|^{-\frac{\alpha+D+2}{2}}}{(2\pi)^{D/2}2^{\frac{\alpha D}{2}}\Gamma_D(\frac{\alpha}{2})} \exp\{-\frac{1}{2}(Tr(\Psi\Sigma^{-1})) + \gamma(\mu-\delta)^T\Sigma^{-1}(\mu-\delta)\}$$

Conjugate Distribution and Conjugate prior

Conjugate Distribution is between prior and posterior

Prior is the conjugate prior of the likelihood function.

2 Fitting model

maximum likelihood

Fitting

$$\begin{aligned}
\hat{\theta} &= argmax_{(\theta)}[Pr(\mathbf{x_{1 \ldots I}}|\theta)] \\
&= argmax_{(\theta)}[\prod_{i=1}^{I} Pr(\mathbf{x_i}|\theta)]
\end{aligned}$$

Predictive Density

Evaluate new data point $\mathbf{x}^*$ under probability distribution $Pr(\mathbf{x}^*|\hat{\theta})$ with best parameter.

maximum a posteriori

Fitting

$$\begin{aligned}
\hat{\theta} &= argmax_{(\theta)}[Pr(\theta|\mathbf{x_{1\ldots I}})] \\
&= argmax_{(\theta)}\left[\frac{Pr(\mathbf{x_{1\ldots I}}|\theta)Pr(\theta)}{Pr(\mathbf{x_{1\ldots I}})}\right] \\
&= argmax_{(\theta)}\left[\frac{\prod_{i=1}^{I} Pr(\mathbf{x_i}|\theta)Pr(\theta)}{Pr(\mathbf{x_{1\ldots I}})}\right] \\
\hat{\theta} &= argmax_{(\theta)}\left[Pr(\mathbf{x_i}|\theta)Pr(\theta)\right]
\end{aligned}$$

Predictive

Evaluate new data point $\mathbf{x}^*$ under probability distribution $Pr(\mathbf{x}^*|\hat{\theta})$ with best parameter.

bayesian approach

Fitting

$$Pr(\theta|\mathbf{x_{1\ldots I}}) = \frac{(\prod_{i=1}^{I} Pr(\mathbf{x_i}|\theta))Pr(\theta)}{Pr(\mathbf{x_{1\ldots I}})}$$

The difference between bayesian approach and MAP is that MAP takes the maximum value, while bayesian approach takes the distribution.

Predictive

$$Pr(\mathbf{x}^*|\mathbf{x_{1\ldots I}}) = \int Pr(\mathbf{x}^*|\theta)Pr(\theta|\mathbf{x_{1\ldots I}})d\theta$$

Confusion: the formula should be $\int Pr(\mathbf{x}^*|\theta, \mathbf{x_{1\ldots I}})Pr(\theta|\mathbf{x_{1\ldots I}})d\theta$. Given the θ, it considers $\mathbf{x_{1\ldots I}}$ and x^* are independent.

Multivariate Normal Distribution

If $\mathbf{x_1}, \mathbf{x_2} \ldots \mathbf{x_n}$ are independent, the covariance matrix would be

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix}$$

Therefore, while $\mathbf{x_1}, \mathbf{x_2} \dots \mathbf{x_n}$ are dependent, the covariance matrix could be decomposed into rotation matrix and diagonal:

$$\Sigma_{full} = \mathbf{R}^T \Sigma'_{diag} \mathbf{R}$$

Marginal Distribution

$$u_i = u_i$$

$$\Sigma_i = \Sigma_{ii}$$

Conditional Distribution

$$u_{i|j} = u_i + \Sigma_{ij} \Sigma_{jj}^{-1} (x_j - u_j)$$

$$\Sigma_{i|j} = \Sigma_{jj} - \Sigma_{ij}^T \Sigma_{ii}^{-1} \Sigma_{ij}$$

Product of two normals

$$\begin{aligned}
\mathrm{Norm}_{\mathbf{x}}[\mathbf{a}, \mathbf{A}] \mathrm{Norm}_{\mathbf{x}}[\mathbf{b}, \mathbf{B}] &= k \cdot \mathrm{Norm}_{\mathbf{x}}[(\mathbf{A^{-1}} + \mathbf{B^{-1}})^{-1}(\mathbf{A^{-1}a} + \mathbf{B^{-1}b}), (\mathbf{A^{-1}} + \mathbf{B^{-1}})^{-1}] \\
k &= \mathrm{Norm}_{\mathbf{a}}[\mathbf{b}, \mathbf{A} + \mathbf{B}]
\end{aligned}$$

change of variables

$$\mathrm{Norm}_{\mathbf{x}}[\mathbf{Ay} + \mathbf{b}, \mathbf{\Sigma}] = k \cdot \mathrm{Norm}_{\mathbf{y}}[\mathbf{A'x} + \mathbf{b'}, \mathbf{\Sigma'}]$$

where

$$\begin{aligned}
\mathbf{A'} &= \Sigma' A^T \Sigma^{-1} \\
b' &= -\Sigma' A^T \Sigma^{-1} b \\
\Sigma &= (A^T \Sigma^{-1} A)^{-1}
\end{aligned}$$

Learning and Inference

The observe measured data, $\mathbf{x}$

Draw inference from it about the state of world, $\mathbf{w}$

If $\mathbf{w}$ is continuous, call this regression.

If $\mathbf{w}$ is discrete, call this classification.

To compute the probability distribution $Pr(\mathbf{w}|\mathbf{x})$, we need: a model(relates visual data $\mathbf{x}$ and $\mathbf{w}$, the relationships depends on parameter θ), a learning algorithm(fits parameter θ from paired training examples $\mathbf{x_i}, \mathbf{w_i}$), an inference algorithm (use model to return $Pr(\mathbf{w}|\mathbf{x})$ given new observed data $\mathbf{x}$)

Types of Model

1. Model contingency of the world on the data $Pr(\mathrm{w}|\mathrm{x})$ (Discriminative models)

1. Choose an appropriate from form for $Pr(\mathbf{w})$
2. Make parameters a function of $\mathbf{x}$
3. Function takes parameters θ that define its shape.

Inference: evaluate $\mathbf{Pr}(\mathrm{w}|\mathrm{x})$

2. Model joint occurrence of the world and data $Pr(\mathrm{x}, \mathrm{w})$ Generative models

1. COncatenate $\mathbf{x}$ and $\mathbf{w}$ to make $\mathbf{z} = [\mathbf{x^T w^T}]$
2. Model of pdf of $\mathbf{z}$
3. Pdf takes parameter θ that define its shape

Inference: compute $Pr(\mathbf{w}|\mathbf{x})$ using Bayes rule.

$$Pr(\mathbf{w}|\mathbf{x}) = \frac{Pr(\mathbf{x}, \mathbf{w})}{Pr(\mathbf{x})} = \frac{Pr(\mathbf{x}, \mathbf{w})}{\int Pr(\mathbf{x}, \mathbf{w}) d\mathbf{w}}$$

3. Model contingency of data on the world $Pr(\mathrm{x}|\mathrm{w})$ (Generative models)

1. Choose an appropriate form for $Pr(\mathbf{x})$
2. Make parameters a function of $\mathbf{w}$
3. Function takes parameter θ that define its shape.

Inference: define prior $Pr(\mathbf{w})$ and then compute $Pr(\mathbf{w}|\mathbf{x})$ using Bayes' rule.

$$Pr(\mathbf{w}|\mathbf{x}) = \frac{Pr(\mathbf{x}|\mathbf{w}) Pr(\mathbf{w})}{\int Pr(\mathbf{x}|\mathbf{w}) \mathbf{Pr}(\mathbf{w}) \mathbf{dw}}$$

Bessel correction

$s^2 = (\frac{n}{n-1})s_n^2$ working this later.

Learning and inference

Mixture of Model

$$Pr(\mathbf{x}|\theta) = \sum_{k=1}^{K} Pr(\mathbf{x}, h = k|\theta)$$

Mixture of Gaussian

$$Pr(\mathbf{x}|\theta) = \sum_{k=1}^{K} \lambda_k \mathrm{Norm}_{\mathbf{x}}[\mu_k, \Sigma_k]$$

Usually, the dimension would be smaller than sample.

Hidden variables

$$\begin{aligned}
Pr(\mathbf{x}) &= \int Pr(\mathbf{x}, \mathbf{h}) d\mathbf{h} \\
Pr(\mathbf{x}|\theta) &= \int Pr(\mathbf{x}, \mathbf{h}|\theta) d\mathbf{h} \\
\hat{\theta} &= \mathrm{argmax}_\theta \left[\sum_{i=1}^{\mathbf{I}} log[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\theta) d\mathbf{h}_i] \right] \\
B[\{q_i(\mathbf{h}_i)\}, \theta] &= \sum_{i=1}^{\mathbf{I}} \int q_i(\mathbf{h}_i) \log[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\theta)}{q_i(\mathbf{h}_i)}] d\mathbf{h}_{1\ldots I} \leq \sum_{i=1}^{\mathbf{I}} log[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\theta) d\mathbf{h}_i]
\end{aligned}$$

Lower bound

Because the log of sum is hard to derivate to 0.

According to Jensen's inequality when $f(x)$ is a convex function:

$$f(\mathbf{E}[\mathbf{X}]) \leq \mathbf{E}[f(\mathbf{X})]$$

For the concave function:

$$f(\mathbf{E}[\mathbf{X}]) \geq \mathbf{E}[f(\mathbf{X})]$$

Therefore the lower bound holds:

$$
\begin{aligned}
\log(\mathbf{E}\left[\frac{Pr(\mathbf{x},\mathbf{h}_i|\theta)}{q(\mathbf{h}_i)}\right]) &\geq \mathbf{E}\left[log(\frac{Pr(\mathbf{x},\mathbf{h}_i|\theta)}{q(\mathbf{h}_i)})\right] \\
\log(\int\left[\frac{Pr(\mathbf{x},\mathbf{h}_i|\theta)}{q(\mathbf{h}_i)}q(\mathbf{h}_i)\right]d\mathbf{h}_i) &\geq \int\left[q(\mathbf{h}_i)log(\frac{Pr(\mathbf{x},\mathbf{h}_i|\theta)}{q(\mathbf{h}_i)})\right]d\mathbf{h}_i \\
\sum_{i=1}^{\mathbf{I}} log[\int Pr(\mathbf{x}_i,\mathbf{h}_i|\theta)d\mathbf{h}_i] &\geq \sum_{i=1}^{\mathbf{I}}\int q_i(\mathbf{h}_i)\log[\frac{Pr(\mathbf{x},\mathbf{h}_i|\theta)}{q_i(\mathbf{h}_i)}]d\mathbf{h}_{1\ldots I}
\end{aligned}
$$

Where $\log$ function is the $f(\mathbf{X})$, and $q(\mathbf{h}_i)$ is $Pr(\mathbf{h}_i|\mathbf{x}_i,\theta^{[t]})$

E-Step

Maximize the bound w.r.t. distribution $q(\mathbf{h}_i)$

$$
\hat{q}_i(\mathbf{h}_i) = Pr(\mathbf{h}_i|\mathbf{x}_i,\theta^{[t]}) = \frac{Pr(\mathbf{x}_i|\mathbf{h}_i,\theta^{[t]})Pr(\mathbf{h}_i|\theta^{[t]})}{Pr(\mathbf{x}_i)}
$$

M-Step

Maximize bound w.r.t parameter θ

$$
\hat{\theta}^{[t+1]} = \text{argmax}_\theta\left[\sum_{i=1}^{I}\int \hat{q}_i(\mathbf{h}_i)\log[Pr(\mathbf{x}_i,\mathbf{h}_i|\theta)]d\mathbf{h}_i\right]
$$

E-step of MoG

$$
\begin{aligned}
Pr(h_i=k|\mathbf{x}_i,\theta^{[t]}) &= \frac{Pr(\mathbf{x}_i|h_i=k,\theta^{[t]})Pr(h_i=k,\theta^{[t]})}{\sum_{j=1}^{K}Pr(\mathbf{x}_i|h_i=j,\theta^{[t]})Pr(h_i=j,\theta^{[t]})} \\
&= \frac{\lambda_k\text{Norm}_{\mathbf{x_i}}[\mu_k,\Sigma_k]}{\sum_{j=1}^{K}\lambda_j\text{Norm}_{\mathbf{x_i}}[\mu_j,\Sigma_j]} \\
&= r_{i,k}
\end{aligned}
$$

M-step of MoG

Take derivative, equal to zero and solve:

$$
\begin{aligned}
\lambda_k^{[t+1]} &= \frac{\sum_{i=1}^{I}r_{i,k}}{\sum_{j=1}^{K}\sum_{i=1}^{I}r_{i,j}} \\
\mu_k^{[t+1]} &= \frac{\sum_{i=1}^{I}r_{i,k}\mathbf{x}_i}{\sum_{i=1}^{I}r_{i,k}} \\
\Sigma_k^{[t+1]} &= \frac{\sum_{i=1}^{I}r_{i,k}(\mathbf{x}_i-\mu_k^{[t+1]})(\mathbf{x}_i-\mu_k^{[t+1]})^T}{\sum_{i=1}^{I}r_{i,k}}
\end{aligned}
$$

Student t-distribution

not willing to write, seems not important

compared to MoG, it is more robustness.

Factor analysis

not willing to write, seems not important

compared to MoG, it is applied when dimension is larger than sample. Or the covariance cannot be invertible.

Regression

Linear Regression

The core idea is to regard the error as the normal distribution.

$$Pr(\mathbf{w}|\mathbf{X}, \theta) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}]$$

Use the maximum likelihood to calculate, then take the derivative, set result to 0 and re-arrange:

$$\begin{aligned}\hat{\phi} &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w} \\ \hat{\sigma}^2 &= \frac{(\mathbf{w} - \mathbf{X}^T\phi)^T(\mathbf{w} - \mathbf{X^T}\phi)}{\mathbf{I}}\end{aligned}$$

Linear Regression in Bayesian

Besides max likelihood, the bayesian model could be applied: Likelihood:

$$Pr(\mathbf{w}|\mathbf{X}, \theta) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}]$$

Prior:

$$Pr(\phi) = \mathrm{Norm}_{\phi}[0, \sigma_p^2\mathbf{I}]$$

Bayes rules:

$$Pr(\phi|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \phi)Pr(\phi|\mathbf{X})}{Pr(\mathbf{w}|\mathbf{X})}$$

In that case, it could be concluded as follow:

$$Pr(\phi|\mathbf{X},\mathbf{w}) = \text{Norm}_\phi[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}]$$

Where $\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}$.

Inference could be calculated as following:

$$\begin{aligned} Pr(w^*|\mathbf{x}^*,\mathbf{X},\mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*,\phi)Pr(\phi|\mathbf{X},\mathbf{w})d\phi \\ &= \text{Norm}_{w^*}[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A^{-1}}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A^{-1}}\mathbf{x}^* + \sigma^2] \end{aligned}$$

where $\mathbf{A^{-1}}$ is hard to calculated when the dimension is large, then directly calculate the $\mathbf{A^{-1}}$.

For the variance fitting, using a marginal distribution to calculate the maximum likelihood:

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\sigma^2}) = \int Pr(\mathbf{w}|\mathbf{X},\phi,\boldsymbol{\sigma^2})Pr(\phi)d\phi = \text{Norm}_\mathbf{w}[0, \sigma_p^2\mathbf{X}^T\mathbf{X} + \boldsymbol{\sigma^2}\mathbf{I}]$$

Gaussian Process Regression

$$Pr(w_i|\mathbf{x}_i,\theta) = \text{Norm}_{w_i}[\phi^T\mathbf{z}_i, \sigma^2]$$

The difference between non-linear regression is using $\mathbf{z}_i$ to substitute $\mathbf{x_i}$. The other steps are similar.

Kernel regression

substitute $\mathbf{Z_i^TZ_i}$ as a kernel $\mathbf{K}[\mathbf{X},\mathbf{X}]$. The advantage is that not waste time on calculating the high dimension $\mathbf{z}$. The specific example could refer to the Gaussian kernel, the $\mathbf{z}$ of Gaussian kernel is infinite dimension. As a kernel, it is calculated fast.

Sparse Linear regression

Perhaps not every dimension of the data $\mathbf{x}$ is informative A sparse solution forces some of the coefficients in ϕ to be zero. The difference between Sparse linear regression and linear regression, here, we applied the t-distribution as the prior as the distribution of ϕ.

The basic idea for this regression is that, t-distribution has a better robustness in data point selection. Then after applying t-distribution as the ϕ distribution, in the fitting phase, fitted ϕ has sparse property.

$$
\begin{aligned}
Pr(\phi) &= \prod_{D}^{d=1} \mathrm{Stud}_{\phi_d}[0,1,\nu] \\
&= \prod_{d=1}^{D} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{\phi_d^2}{\nu})^{-(\nu+1)/2}
\end{aligned}
$$

For every single t-distribution, it could be regarded as a mixture of Gaussian distribution. Therefore, it could be expressed as follow (hidden variable):

$$
\begin{aligned}
Pr(\phi) &= \prod_{d=1}^{D} \int \mathrm{Norm}_{\phi_d}[0,1/h_d]\mathrm{Gam}_{h_d}[\nu/2,\nu/2]dh_d \\
&= \int \mathrm{Norm}_{\phi_d}[0,\mathbf{H}^{-1}] \prod_{d=1}^{D} \mathrm{Gam}_{h_d}[\nu/2,\nu/2]d\mathbf{H}
\end{aligned}
$$

Then, according to bayesian:

$$
Pr(\mathbf{w}|\mathbf{X},\sigma^{\mathbf{2}}) \approx \max_{\mathbf{H}}[\mathrm{Norm}_{\mathbf{w}}[0,\mathbf{X^T H^{-1} X}+\sigma^{\mathbf{2}}\mathbf{I}] \prod_{d=1}^{D} \mathrm{Gam}_{h_d}[\nu/2,\nu/2]]
$$

The specific result could refer to ppt08 page 44.

However, it is hard to handle high dimension.

Dual Linear Regression

This model could be regarded as SVM regression in bayesian framework.

In linear SVR, the regressor is:

$$
y = \sum_{i=1}^{N}(a_i - a_i^*) < x_i, x > +b
$$

where a_i is the upper bound penalty factor Lagrange multiplier, and a_i^* is lower penalty factor Lagrange multiplier. Then $(a_i - a_i^*)$ could be regarded as a new coefficient. Then introduce the dual linear regression. The specific SVR tutorial URL could be found in http://kernelsvm.tripod.com/. The idea is that ϕ could be represented as

$$
\phi = \mathbf{X}\psi = \sum_{i=1}^{N}(a_i - a_i^*)x_i
$$

Then dual linear regression could be represented as :

$$
Pr(\mathbf{w}|\mathbf{X},\theta) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\mathbf{X}\psi, \sigma^2\mathbf{I}]
$$

The fitting and inferencing are the same as linear regression above.

Relevance Vector Machine

The idea is to combine dual regression and sparsity.

$$Pr(\mathbf{w}|\mathbf{X},\theta) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X^T X}\psi, \sigma^2\mathbf{I}]$$

$$Pr(\psi) = \prod_{i=1}^{I} \mathrm{Stud}_{\psi_i}[0, 1, \nu]$$

Classification

Logistic Regression

$$Pr(w|\phi_0, \phi, \mathbf{x}) = \mathrm{Bern}_w[\sigma(\phi^T\mathbf{x})]$$

If use the maximum likelihood to learning, the gradient would be

$$\frac{\partial L}{\partial \phi} = -\sum_{i=1}^{I}(\frac{1}{1+\exp[-\phi^T\mathbf{x}_i]} - w_i)\mathbf{x}_i = -\sum_{i=1}^{I}(\mathrm{sig}[a_i] - w_i)\mathbf{x}_i$$

we cannot get the expression for ϕ in term of x and w. Therefore, the goal is to optimize that

$$\hat{\theta} = \mathrm{argmin}_{\theta}[f[\theta]]$$

If a function is convex, then it has only a single minimum.

Gradient Based Optimization

1. Choose a search direction s based on the local properties of the function.
2. Perform an intensive search along the chosen direction. This is called line search:

$$\hat{\lambda} = \mathrm{argmin}_{\lambda}[f[\theta^{[t]} + \lambda\mathbf{s}]]$$

Then set

$$\theta^{[t+1]} = \theta^{[t]} + \hat{\lambda}\mathbf{s}$$

In order to solve the not compute gradient problem, there is the solution that:

$$\frac{\partial f}{\partial \theta_j} \approx \frac{f[\theta + a\mathbf{e}_j] - f[\theta]}{a}$$

where $\mathbf{e_j}$ is the unit vector in the j^{th} direction.

Newton's method

$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \lambda(\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2})^{-1}\frac{\partial f}{\partial \boldsymbol{\theta}}$

Line Search

It is the similar to Golden-section search. Choose a range, and split it into 3 range(a,b,c,d). $a, d = f(b) > f(c) ? b, d : a, c$

Bayesian Logistic Regression

$$Pr(\phi|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \phi)Pr(\phi)}{Pr(\mathbf{w}|\mathbf{X})}$$

where $Pr(\phi) = \text{Norm}_\phi[0, \sigma_\text{p}^2\mathbf{I}]$.

Laplace Approximation

Set mean to MAP estimate
Set covariance to match that at MAP estimate.

$$Pr(\phi|\mathbf{X}, \mathbf{w}) \approx q(\phi) = \text{Norm}_\phi[\boldsymbol{\mu}, \Sigma]$$

where $\boldsymbol{\mu} = \hat{\phi}$ and $\mathbf{\Sigma} = -\left(\frac{\partial^2 L}{\partial \phi^2}\right)^{-1}|_{\phi=\hat{\phi}}$

Inference

$$\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \phi)Pr(\phi|\mathbf{X}, \mathbf{w})d\phi \\
&\approx \int Pr(w^*|\mathbf{x}^*, \phi)q(\phi)d\phi \\
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &\approx \int Pr(w^*|a)Pr(a)da \\
&\approx \frac{1}{1 + \exp[-\mu_a/\sqrt{1 + \pi\sigma_a^2/8}]} \\
Pr(a) = Pr(\phi^T\mathbf{x}^*) &= \text{Norm}_a[\mathbf{u}^T\mathbf{x}^*, \mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}] \\
&= \text{Norm}_a[\mu_a, \sigma_a^2]
\end{aligned}$$

Non-linear logistic regression

Apply non-linear transformation:

$$\mathbf{z} = \mathbf{f}[\mathbf{x}]$$

Build model as usual

$$Pr(w = 1|\mathbf{x}, \phi) = \text{Bern}_w[\text{sig}[\phi^T \mathbf{z}]]$$

Gaussian process classification

Combine Bayesian logistic regression and non-linear logistic regression in dual form.
The likelihood is:

$$Pr(\mathbf{w}|\mathbf{X}, \psi) = \prod_{i=1}^{I} Bern_{w_i}[\text{sig}[a_i]] = \prod_{i=1}^{I} Bern_{w_i}[\text{sig}[\psi^T \mathbf{X}^T x_i]]$$

Relevance vector classification

It is the same as regression model that use t-distribution as priori of dual form coefficient.

Incremental fitting

In the previous models, we wrote as following:

$$a_i = \phi^T \mathbf{z}_i = \phi^T \mathbf{f}[\mathbf{x}]_i$$

Now write:

$$a_i = \phi_0 + \sum_{k=1}^{K} \phi_k f[\mathbf{x}_i, \xi_k]$$

The derivate would be

$$\frac{\partial L}{\partial \theta} = -\sum_{i=1}^{I} (w_i - \text{sig}[a_i]) \frac{\partial a_i}{\partial \theta}$$

Branching logistic regression

$$a_i = (1 - g[\mathbf{x}_i, \boldsymbol{w}]) \phi_0^T \mathbf{x}_i + g[\mathbf{x}_i, \boldsymbol{w}] \phi_1^T \mathbf{x}_i$$

where $g[*, *]$ is a gating function.