Rui Qiu #999292509

200

STA 414 中国作家协会
Assignment 2.

#1. Solution: posterior $\propto$ likelihood $\cdot$ prior

we know likelihood: $p(t|X,w,\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$

prior: $p(w)=\mathcal{N}(w|m_0,S_0)$

so $p(w|t)\propto\left\{\prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})\right\}\mathcal{N}(w|m_0,S_0)$

posterior

ignore constant terms

$$\propto \exp\left(-\frac{\beta}{2}(t-\Phi w)^T(t-\Phi w)\right)\cdot\exp\left(-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right)$$

$$=\exp\left(-\frac{1}{2}\left[\beta t^Tt-\beta w^T\Phi^Tt-\beta t^T\Phi w+w^T\beta\Phi^T\Phi w + w^TS_0^{-1}w-w^TS_0^{-1}m_0-m_0^TS_0^{-1}w+m_0^TS_0^{-1}m_0\right]\right)$$

$$=\exp\left(-\frac{1}{2}\left[w^T(\beta\Phi^T\Phi+S_0^{-1})w-\beta w^T\Phi^Tt-\beta t^T\Phi w+\beta t^Tt - m_0^TS_0^{-1}w-w^TS_0^{-1}m_0+m_0^TS_0^{-1}m_0\right]\right)$$

$$=\exp\left(-\frac{1}{2}\left[w^T(\beta\Phi^T\Phi+S_0^{-1})w-(\beta\Phi^Tt+S_0^{-1}m_0)^Tw - w^T(\beta\Phi^Tt+S_0^{-1}m_0) + \beta t^Tt + m_0^TS_0^{-1}m_0\right]\right)$$

10

$$=\exp\left(-\frac{1}{2}\left[w^TS_n^{-1}w \cancel{-m_N^Tw-w^Tm_N} -(m_N\cdot S_N^{-1})^Tw - w^T(m_N\cdot S_N^{-1}) + \beta t^Tt+m_0^TS_0^{-1}m_0\right]\right) \quad \text{by } (*)$$

$$\cancel{=\exp\left(-\frac{1}{2}[w^TS_n^{-1}w-m_0^TS_0^{-1}w+\beta t^T\Phi w-w^Tm_N\right)}$$

$$=\exp\left(-\frac{1}{2}\left[w^TS_n^{-1}w-(m_N S_N^{-1})^Tw-w^T(m_N S_N^{-1})+m_N^TS_N^{-1}m_N\right]\right)\cdot \exp\left(-\frac{1}{2}\left[\beta t^Tt+m_0^TS_0^{-1}m_0-m_N^TS_N^{-1}m_N\right]\right)$$

"borrow this term from nowhere."

Therefore, the posterior distribution is given by a multivariate Gaussian distribution ~~which only depends on m_N~~ (In details, the first exp is unnormalized Gaussian the second exp is the inverse of normalization factor term.) which needs to be eliminated.

(*). b/c $m_N=S_N(S_0^{-1}m_0+\beta\Phi^Tt)$

$S_N^{-1}=S_0^{-1}+\beta\Phi^T\Phi$

$(m_N\cdot S_N^{-1})^T=(S_0^{-1}m_0+\beta\Phi^Tt)^T$

$=m_0^TS_0^{-1}+\beta t^T\Phi$

#2. Solution:

As $\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$ (***), in order to show some conclusion about $\sigma_{N+1}^2(x)$, we need problem #1 to show S_{N+1} (S_N is shown in #1)

- prior in #1. $p(w) = N(w \mid m_N, S_N)$
- likelihood $p(t_{N+1} \mid x_{N+1}, w) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\beta}{2}(t_{N+1} - w^T \underbrace{\phi_{N+1}}_{\text{this is } \phi(x_{N+1})})^2\right)$

so the posterior, again:

$$p(w \mid t_{N+1}, x_{N+1}, m_N, S_N) \propto \exp\left(-\frac{\beta}{2}(t_{N+1} - w^T\phi_{N+1})^2 - \frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)\right)$$

$$\propto w^T S_N^{-1} w - 2w^T S_N^{-1} m_N + \beta w^T \phi_{N+1}\phi_{N+1}^T w - 2\beta w^T \phi_{N+1} t_{N+1} + \text{const}$$

$$= w^T(S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T)w - 2w^T(\beta\phi_{N+1}t_{N+1} + S_N^{-1}m_N) + \text{const}$$

Note const ~~is~~ are term(s) with no w inside.

~~This is what we want $= w^T S_{N+1}^{-1} w - 2w^T S_{N+1}^{-1} S_{N+1} + \text{const}$~~

so $\underbrace{S_{N+1}^{-1} = S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T}_{(*)}$, $m_{N+1} = S_{N+1}(\beta\phi_{N+1}t_{N+1} + S_N^{-1}m_N)$

as $\beta\phi_{N+1}t_{N+1} + S_N^{-1}m_N = m_{N+1} \cdot S_{N+1}^{-1}$.

Now, using matrix identity $(M + vv^T)^{-1} = M^{-1} - \frac{(M^{-1}v)(v^TM^{-1})}{1 + v^TM^{-1}v}$ (**)

& (*) we get:

$$S_{N+1} = (S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T)^{-1} = S_N - \frac{(S_N\beta^{\frac{1}{2}}\phi_{N+1})(\phi_{N+1}^T S_N \beta^{\frac{1}{2}})}{1 + \phi_{N+1}^T S_N \phi_{N+1} \cdot \beta}$$

10

$$= S_N - \frac{\beta S_N \phi_{N+1}\phi_{N+1}^T S_N}{1 + \beta\phi_{N+1}^T S_N \phi_{N+1}}$$

$$(***) \Rightarrow \sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi(x)^T S_{N+1}\phi(x) = \frac{1}{\beta} + \phi(x)^T\left(S_N - \frac{\beta S_N\phi_{N+1}\phi_{N+1}^T S_N}{1 + \beta\phi_{N+1}^T S_N \phi_{N+1}}\right)\phi(x)$$

$$= \sigma_N^2(x) - \phi(x)^T\left(\frac{\beta S_N \phi_{N+1}\phi_{N+1}^T S_N}{1 + \beta\phi_{N+1}^T S_N \phi_{N+1}}\right)\phi(x)$$

As ~~S_N is positive definite,~~ $\phi(x)^T \beta S_N \phi_N \phi_{N+1}^T S_N \phi(x) \geq 0$

and $1 + \beta\phi_{N+1}^T S_N \phi_{N+1} > 0$

so $\sigma_{N+1}^2(x) - \sigma_N^2(x) \leq 0$.

中国作家协会

#3 Solution:

① Maximum likelihood solution for the prior probabilities:

Take logarithm:

$$\ln p(\{\phi_n, t_n\}|\{\pi_k\}) = \ln\left(\prod_{n=1}^{N}\prod_{k=1}^{K}(p(\phi_n|C_k)\pi_k)^{t_{nk}}\right), \quad \phi_n = \phi(x_n)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\ln p(\phi_n|C_k) + \ln\pi_k) \qquad (*)$$

In order to maximize this logarithm, it is the same to maximize the following with Lagrange ~~multi~~ multiplier λ:

$$\ln p(\phi_n, t_n|\pi_k) + \lambda(\textstyle\sum_{k=1}^{k}\pi_k - 1) \qquad \text{as } \sum_{k=1}^{k}\pi_k = 1 \text{ is a "constraint"}$$

Take derivative w.r.t. π_k:

$$\sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda = 0$$

$$\sum_{n=1}^{N} t_{nk} = -\pi_k\lambda = N_k \quad \text{by definition}$$

$$\lambda = \frac{\sum_{n=1}^{N} t_{nk}}{-\pi_k} = \frac{N_k}{-\pi_k} = -N$$

so $\pi_k = \dfrac{N_k}{N}$ as desired.

10

② Maximum likelihood solution for the mean of the Gaussian dist'n for class C_k:

Since $p(x|C_k) = \mathcal{N}(x|\mu_k, \Sigma)$ and $(*)$, we combine them together:

$$\ln p(\{\phi_n, t_n\}|\{\pi_k\}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\ln|\Sigma| + (\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k)) + \text{const.} \quad (**)$$

where const term is independent of μ_k & Σ.

Set $(**)$'s derivative to zero, according to $\frac{\partial}{\partial x}(x^Ta) = \frac{\partial}{\partial x}(a^Tx) = a$

$$\left.\begin{array}{l}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\Sigma^{-1}(\phi_n - \mu_k) = 0 \\ \text{recall } \sum_{n=1}^{N} t_{nk} = N_k\end{array}\right\} \Rightarrow N_k \cdot \mu_k = \sum_{n=1}^{N} t_{nk}x_n = \sum_{n=1}^{N} t_{nk}(\Sigma^{-1}\phi_n)$$

Therefore $\mu_k = \dfrac{1}{N_k}\sum_{n=1}^{N} t_{nk}x_n$ as desired.

③ Maximum likelihood solution for the shared covariance matrix.

(see next page)

We want to write $(**)$, take derivative of it again, but this time w.r.t. Σ^{-1}.

Recall: $\frac{\partial}{\partial A} \ln|A| = (A^{-1})^T$

$\frac{\partial}{\partial A} Tr(AB) = B^T$

And by $(**)$: $\ln p(\{\phi_n, t_n\}|\{\pi_k\}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\ln|\Sigma| + (\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k))$

$= -\frac{1}{2} m \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\ln|\Sigma| + Tr(\Sigma^{-1}(\phi_n - \mu_k)(\phi_n - \mu_k)^T))$

for some constant m. $(***)$

Take derivative w.r.t Σ^{-1} (NOT Σ!)

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\Sigma - (\phi_n - \mu_k)(\phi_n - \mu_k)^T) = 0$$

Again, use $\sum_{n=1}^{N} t_{nk} = N_k$

So $S_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$. similarly like last part. ∎

#4. Solution:

$w^T x = 0$ is the decision boundary:

$$w^T \phi_n = \begin{cases} \geq 0 & \text{if } t_n = 1 \\ < 0 & \text{o.w.} \end{cases} \quad \text{where } \phi_n = \phi(x_n)$$

Also note that:

The cross-entropy error function:

$$E(w) = -\ln p(t|w) = -\sum_{n=1}^{N} (t_n \ln y_n + (1-t_n)\ln(1-y_n))$$

10

is minimized (since negative) when $y_n = t_n$

$$b/c \quad \frac{t_n}{y_n} + \frac{1-t_n}{1-y_n}\cdot(-1) = \frac{t_n}{y_n} - \frac{1-t_n}{1-y_n} = \frac{(1-y_n)t_n - y_n(1-t_n)}{y_n(1-y_n)} = \frac{t_n - y_n}{y_n(1-y_n)} = 0$$

$$\Rightarrow t_n = y_n = \sigma(w^T\phi) \qquad (*)$$

so $|w| \to \infty$ is the only condition that makes $(*)$ happen.

#5. Solution:

Problem is a binary classification: for each point,

① $t_n = 1$, it has prob. $p(t_n=1|\phi_n)$

② $t_n = 0$, it has prob. $p(t_n=0|\phi_n) = 1 - p(t_n=1|\phi_n)$ where $\phi(x_n) = \phi_n$

We know, each x_n, ~~have~~ we have a value of π_n representing the probability that $t_n = 1$, so ①② $\Rightarrow$

$$\pi_n p(t_n=1|\phi_n) + (1-\pi_n)(1-p(t_n=1|\phi_n))$$

Take logarithm:

$$\sum_{n=1}^{N} \{\pi_n \ln p(t_n=1|\phi_n) + (1-\pi_n) \ln(1-p(t_n=1|\phi_n))\}$$

This is the desired log-likelihood function.

10