WORKSHEET 6

TRAINING DATA $\{t_n, \underline{x}_n\}_{n=1}^{N}$

$\underline{x}_n = (x_{n0}, \ldots, x_{n,D-1})^T$, $X = (\underline{x}_1, \ldots, \underline{x}_N)^T$

$t_n \in \{1, \ldots, C\}$, $\underline{t} = (t_1, \ldots, t_N)^T$

$\therefore$ FOR EACH CLASS c WE HAVE TRAINING DATA

$\{t_n, \underline{x}_n\}_{n=1}^{N_c}$, $t_n = c$

$X^c = (\underline{x}_1, \ldots, \underline{x}_{N_c})^T$, $\underline{t}^c = (t_1, \ldots, t_{N_c})^T$

WANT TO DEVELOP A PROBABILISTIC CLASSIFIER WHICH FOR SOME NEW DATA $\underline{x}_*$ PROVIDES

$$p(t_* = c \mid \underline{x}_*, X, \underline{t})$$

A BAYES CLASSIFIER WILL PRODUCE

$$p(t_* = c \mid \underline{x}_*, X, \underline{t}) = \frac{p(\underline{x}_* \mid t_* = c, X, \underline{t})\, p(t_* = c \mid X, \underline{t})}{\sum_{c'=1}^{C} p(\underline{x}_* \mid t_* = c', X, \underline{t})\, p(t_* = c \mid X, \underline{t})}$$

$p(\underline{x}_* \mid t_* = c, X, \underline{t})$ - Likelihood term - Distribution of $\underline{x}_*$ specific to class c

$p(t_* = c \mid X, \underline{t})$ - prior term - Beliefs about frequency of class c.

PROBLEM 1

GAUSSIAN LIKELIHOOD / CLASS CONDITIONALS

ASSUME $\underline{x}_* \mid t_* = c, X, t \sim \mathcal{N}(\mu_c, \Sigma_c)$

NEED TO FIND $\mu_c, \Sigma_c \mid X^c, t^c$

ASSUME $\Sigma_c = I$

FIND $\mu_c \mid X^c$

BAYES RULE : $p(\mu_c \mid X_c) = \dfrac{p(X^c \mid \mu_c)\, p(\mu_c)}{p(X^c)}$

NOTE: ASSUMING THAT $\underline{x}_i \mid \mu_c, \Sigma_c$ is iid

Then $p(X^c \mid \mu_c, \Sigma_c = I) = p(\underline{x}_1, \ldots, \underline{x}_{N_c} \mid \mu_c, \Sigma_c = I)$

$$= \prod_{n=1}^{N_c} \cancel{\mathcal{N}} \; p(x_n \mid \mu_c, \Sigma_c)$$

$$= \prod_{n=1}^{N_c} \mathcal{N}(\mu_c, I)$$

ASSUME $\mu_c \sim \mathcal{N}(\mu_0, \Sigma_0)$

THEN : $p(\mu_c \mid X_c) \propto \mathcal{N}(\mu_0, \Sigma_0) \prod_{n=1}^{N_c} \mathcal{N}(\mu_c, I)$

AS THIS is A PRODUCT OF GAUSSIAN DISTRIBUTIONS

$$\mu_c \mid X_c, \mu_0, \Sigma_0, \Sigma_c = I \sim \mathcal{N}(m, S)$$

AND $\mathcal{N}(m, S) \propto \exp\left\{ -\tfrac{1}{2} \left(\mu_c^T S^{-1} \mu_c - 2\, m^T S^{-1} \mu_c \right) \right\}$

AND SO: $p(\mu_c | X_c, \Sigma_c = I, \mu_0, \Sigma_0)$

$$\propto \mathcal{N}(\mu_0, \Sigma_0) \prod_{n=1}^{N_c} \mathcal{N}(\mu_c, I)$$

$$\propto \exp\left\{-\frac{1}{2}\left[(\mu_c-\mu_0)^T \Sigma_0^{-1}(\mu_c-\mu_0)\right]\right\} \exp\left\{-\frac{1}{2}\left[\sum_{n=1}^{N_c}(x_n-\mu_c)^T(x_n-\mu_c)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\mu_c^T \Sigma_0^{-1}\mu_c - 2\mu_0^T\Sigma_0^{-1}\mu_c + N_c\,\mu_c^T\mu_c - 2\left(\sum_{n=1}^{N_c} x_n^T\right)\mu_c\right)\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\mu_c^T\left(\Sigma_0^{-1} + N_c I\right)\mu_c - 2\left(\mu_0^T\Sigma_0^{-1} + \sum_{n=1}^{N_c} x_n^T\right)\mu_c\right)\right\}$$

$$\Rightarrow \quad S = \left(\Sigma_0^{-1} + N_c I\right)^{-1}$$

$$m^T S^{-1} = \left(\mu_0^T \Sigma_0^{-1} + \sum_{n=1}^{N_c} \underline{x}_n^T\right)$$

$$m = S\left(\Sigma_0^{-1}\mu_0 + \sum_{n=1}^{N_c} \underline{x}_n\right)$$

PROBLEM 2

WE ARE ONLY REALLY INTERESTED IN μ_c AND Σ_c AS A MEANS TO CALCULATE THE CLASS CONDITIONAL LIKELIHOOD

$$p(x_* | t_* = c, X, t)$$

HOWEVER, BECAUSE WE HAVE CALCULATED A

POSTERIOR DISTRIBUTION FOR μ_c, THIS IMPLIES WE NOW HAVE A DISTRIBUTION FOR THE CLASS CONDITIONAL LIKELIHOOD.

WE REALLY WANT TO BE ABLE TO PLUG A SCALAR VALUE ~~FOR~~ $p(\underline{x}_* | t_* = c, X, t)$ INTO OUR BAYES CLASSIFIER.

ONE GOOD WAY OF SUMMARISING A DISTRIBUTION WITH A SCALAR IS TO TAKE IT'S EXPECTATION.

THUS WE WANT TO CALCULATE THE EXPECTATION OF THE CLASS CONDITIONAL LIKELIHOOD WITH RESPECT TO THE POSTERIOR DISTRIBUTION FOR μ_c. THAT IS WE LET

$$p(\underline{x}_* | t_* = c, X, t) = \mathbb{E}_{p(\mu_c | X^c)}\left[p(\underline{x}_* | \mu_c, \Sigma_c = I) \right]$$

$$= \int_{-\infty}^{\infty} p(\underline{x}_* | \mu_c, \Sigma_c)\, p(\mu_c | X^c)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(\underline{x}_* | \mu_c, \Sigma_c)\, \mathcal{N}(\mu_c | m, S)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(\mu_c | \underline{x}_*, \Sigma_c)\, \mathcal{N}(\mu_c | m, S)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(\underline{x}_* | m, \Sigma_c + S)\, \mathcal{N}(\mu_c | m', S')\, d\mu_c$$

$$= \mathcal{N}(\underline{x}_* | m, S) \underbrace{\int_{-\infty}^{\infty} \mathcal{N}(\mu_c | m', S')\, d\mu_c}_{= 1 \text{ as pdf}}$$

$$= \mathcal{N}(\underline{x}_* | m, S)$$

where the above operations rely on standard results for

PROBLEM 3

RATHER THAN PERFORMING A FULL BAYESIAN ANALYSIS BY CALCULATING A POSTERIOR DISTRIBUTION FOR μ_c AND Σ_c, AN ALTERNATIVE APPROACH IS SIMPLY TO ESTIMATE μ_c AND Σ_c BY MAXIMUM LIKELIHOOD

SO FOR CLASS c with TRAINING DATA

$$\left\{ t_n, \underline{x}_n \right\}_{n=1}^{N_c} \quad , \quad t_n = c \quad , \quad X^c = (x_1, \dots, x_{N_c})^T$$

WE ASSUME $\quad \underline{x}_n \mid t_n = c \overset{iid}{\sim} \mathcal{N}(\mu_c, \Sigma_c)$

THEREFORE THE LIKELIHOOD OF THE DATA

$$\begin{aligned}
\mathcal{L} &= p(X^c \mid t = c, \mu_c, \Sigma_c) \\
&= p(\underline{x}_1, \dots, \underline{x}_{N_c} \mid t = c, \mu_c, \Sigma_c) \\
&= \prod_{n=1}^{N_c} p(\underline{x}_n \mid t = c, \mu_c, \Sigma_c) \\
&= \prod_{n=1}^{N_c} \mathcal{N}(\underline{x}_n \mid \mu_c, \Sigma_c) \\
&= \prod_{n=1}^{N_c} \left(\frac{1}{|2\pi\Sigma_c|} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\underline{x}_n - \mu_c)^T \Sigma_c^{-1} (\underline{x}_n - \mu_c) \right\} \\
&= \left(|2\pi \Sigma_c| \right)^{-\frac{N_c}{2}} \exp\left\{ -\frac{1}{2} \sum_{n=1}^{N_c} (\underline{x}_n - \mu_c)^T \Sigma_c^{-1} (\underline{x}_n - \mu_c) \right\}
\end{aligned}$$

THE LOG LIKELIHOOD IS THEN

$$\ell = \log \mathcal{L}$$

$$= -\frac{N_c}{2} \log |2\pi \Sigma_c| - \frac{1}{2} \sum_{n=1}^{N_c} (\underline{x}_n - \mu_c)^T \Sigma_c^{-1} (x_n - \mu_c)$$

$$= -\frac{N_c \cdot D}{2} \log 2\pi - \frac{N_c}{2} \log |\Sigma_c| - \frac{1}{2} \sum_{n=1}^{N_c} (x_n - \mu_c)^T \Sigma_c^{-1} (x_n - \mu_c)$$

FIND ML ESTIMATOR FOR μ_c by finding $\frac{\partial \ell}{\partial \mu_c}$ and setting the result to 0

$$\frac{\partial \ell}{\partial \mu_c} = -\frac{1}{2} \sum_{n=1}^{N_c} 2(-1) \Sigma_c^{-1} (x_n - \mu_c) = 0$$

$$\Rightarrow \sum_{n=1}^{N_c} (x_n - \mu_c) = 0$$

$$\Rightarrow \hat{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \underline{x}_n$$

ESTIMATOR FOR Σ_c

$$\frac{\partial \ell}{\partial \Sigma_c} = -\frac{N_c}{2} \Sigma_c^{-1} + \frac{1}{2} \sum_{n=1}^{N_c} \Sigma_c^{-1} (\underline{x}_n - \mu_c)(x_n - \mu_c)^T \Sigma_c^{-1} = 0$$

$$\Rightarrow \sum_{n=1}^{N_c} (\underline{x}_n - \mu_c)(x_n - \mu_c)^T \Sigma_c^{-1} = N_c$$

$$\Rightarrow \hat{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (x_n - \mu_c)(x_n - \mu_c)^T$$

See 57 & 61 of the Matrix Cookbook

A SECOND DERIVATIVE TEST WILL VERIFY THAT THE LIKELIHOOD HAS BEEN MAXIMISED

APPENDIX

$$p(x_* \mid t_* = c, X, \underline{t}) = \mathbb{E}_{p(\mu_c \mid X^c)}\left[p(x_* \mid \mu_c, \Sigma_c) \right]$$

$$= \int_{-\infty}^{\infty} p(x_* \mid \mu_c, \Sigma_c)\, p(\mu_c \mid X^c)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(x_* \mid \mu_c, \Sigma_c)\, \mathcal{N}(\mu_c \mid m, S)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(\mu_c \mid x_*, \Sigma_c)\, \mathcal{N}(\mu_c \mid m, S)\, d\mu_c$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{|2\pi \Sigma_c|}\right)^{1/2} \exp\left\{-\frac{1}{2}\left((\mu_c - x_*)^T \Sigma_c^{-1} (\mu_c - x_*)\right)\right\} \cdot$$

$$\left(\frac{1}{|2\pi S|}\right)^{1/2} \exp\left\{-\frac{1}{2}\left((\mu_c - m)^T S^{-1} (\mu_c - m)\right)\right\} d\mu_c$$

See Eq. 346 of Matrix cookbook.

$$= \left(\frac{1}{|2\pi \Sigma_c|}\right)^{1/2} \left(\frac{1}{|2\pi S|}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(\mu_c - x_*)^T \Sigma_c^{-1} (\mu_c - x_*)\right.$$

$$\left. -\frac{1}{2}(\mu_c - m)^T S^{-1} (\mu_c - m)\right\} d\mu_c$$

$$= \left(\frac{1}{|2\pi \Sigma_c|}\right)^{1/2} \left(\frac{1}{|2\pi S|}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\mu_c - (\Sigma_c^{-1} + S^{-1})^{-1}(\Sigma_c^{-1} x_* + S^{-1} m)\right)^T\right.$$

$$\left.(\Sigma_c^{-1} + S^{-1})\left(\mu_c - (\Sigma_c^{-1} + S^{-1})^{-1}(\Sigma_c^{-1} x_* + S^{-1} m)\right)\right\}$$

$$\exp\{C\}\, d\mu_c$$

where C is independent of μ_c. See 358 - 364,

(12)

$$\Rightarrow \left(\frac{1}{|2\pi \Sigma_c|}\right)^{\frac{1}{2}} \left(\frac{1}{|2\pi S|}\right)^{\frac{1}{2}} \exp\{C\} \left(|2\pi(\Sigma_c^{-1}+S^{-1})^{-1}|\right)^{\frac{1}{2}}$$

where $\left(|2\pi(\Sigma_c+S)^{-1}|\right)^{\frac{1}{2}}$ is the normalising constant for a $\mathcal{N}\left((\Sigma_c^{-1}+S^{-1})^{-1}(\Sigma_c^{-1}x_* + S m), (\Sigma_c+S)^{-1}\right)$ distribution

Note:

$$C = \frac{1}{2}\left((x_*^T\Sigma_c^{-1} + m S^{-1})(\Sigma_c^{-1} + S^{-1})^{-1}(\Sigma_c^{-1}x_* + S^{-1}m)\right) - \frac{1}{2}\left(x_*^T\Sigma_c^{-1}x_* + m^T S^{-1} m\right)$$

$$= \frac{1}{2}\Big(x_*^T\Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}\Sigma_c^{-1}x_* + 2x_*^T\Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}S^{-1}m$$

$$+ mS^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}S^{-1}m - x_*^T\Sigma_c^{-1}x_* - m^T S^{-1}m\Big)$$

$$= \frac{1}{2}\Big(x_*^T\left[\Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}\Sigma_c^{-1} - \Sigma_c^{-1}\right]x_* + 2x_*^T\Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}S^{-1}m$$

$$+ m^T\left[S^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}S^{-1} - S^{-1}\right]m\Big)$$

THEN AS: $\Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}\Sigma_c^{-1} - \Sigma_c^{-1}$

$$= \Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}\left(\Sigma_c^{-1} - (\Sigma_c^{-1}+S^{-1})\right)$$

$$= \Sigma_c^{-1}(\Sigma_c^{-1}+S^{-1})^{-1}(-S^{-1})$$

$$= -\left(\Sigma_c(\Sigma_c^{-1}+S^{-1})S\right)^{-1}$$

$$= (\Sigma_c+S)^{-1}$$

$$\Rightarrow \quad C = -\frac{1}{2}\Big(x_*^T (\Sigma_c + S)^{-1} x_* - 2 x_*^T (\Sigma_c + S)^{-1} m$$

$$+ m^T (\Sigma_c + S)^{-1} m \Big)$$

$$= -\frac{1}{2}\Big((x_*^T - m^T)(\Sigma + S)^{-1}(x_* - m) \Big)$$

$$\Rightarrow \left(\frac{1}{|2\pi \Sigma_c|}\right)^{\frac{1}{2}} \left(\frac{1}{|2\pi S|}\right)^{\frac{1}{2}} \left(|2\pi (\Sigma_c^{-1} + S^{-1})^{-1}|\right)^{\frac{1}{2}} \cdot$$

$$\exp\left\{ -\frac{1}{2} (x_* - m)^T (\Sigma + S)^{-1} (x_* - m) \right\}$$

CONSIDER $\left(\frac{1}{|2\pi \Sigma_c|}\right)^{\frac{1}{2}} \left(\frac{1}{|2\pi S|}\right)^{\frac{1}{2}} \left(|2\pi (\Sigma_c^{-1} + S^{-1})^{-1}|\right)^{\frac{1}{2}}$

$$= |2\pi \Sigma_c|^{-\frac{1}{2}} \, |2\pi S|^{-\frac{1}{2}} \, |2\pi (\Sigma_c^{-1} + S^{-1})^{-1}|^{\frac{1}{2}}$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c|^{-\frac{1}{2}} (2\pi)^{-\frac{D}{2}} |S|^{-\frac{1}{2}} (2\pi)^{\frac{D}{2}} |(\Sigma_c^{-1} + S^{-1})^{-1}|^{\frac{1}{2}}$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c|^{-\frac{1}{2}} |(\Sigma_c^{-1} + S^{-1})^{-1}|^{\frac{1}{2}} |S|^{-\frac{1}{2}}$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c^{-\frac{1}{2}}| \, |(\Sigma_c^{-1} + S^{-1})^{-\frac{1}{2}}| \, |S^{-\frac{1}{2}}|$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c^{-\frac{1}{2}} (\Sigma_c^{-1} + S^{-1})^{-\frac{1}{2}} S^{-\frac{1}{2}}|$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c (\Sigma_c^{-1} + S^{-1}) S|^{-\frac{1}{2}}$$

$$= (2\pi)^{-\frac{D}{2}} |\Sigma_c + S|^{-\frac{1}{2}}$$

$$= |2\pi (\Sigma_c + S)|^{-\frac{1}{2}}$$

See 18-24 of Matrix Cookbook.

$$\Rightarrow \left(\frac{1}{|2\pi(\Sigma_c + S)|} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (x_* - m)^T (\Sigma_c + S)^{-1} (x_* - m) \right\}$$

$$= \mathcal{N}(x_* \mid m, \Sigma_c + S)$$

And so the likelihood at $\underline{x}_*$ is given by the $\mathcal{N}(m, \Sigma_c + S)$ pdf, providing a scalar value for any $\underline{x}_*$