Back Propagation:

$$L(x) = L(f(h(g(x)))$$

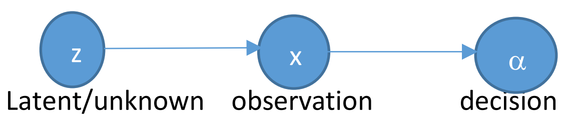
$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial x} = \lambda_g \frac{\partial g}{\partial x}$$

$$\lambda_f = \frac{\partial L}{\partial f}; \lambda_h = \frac{\partial L}{\partial h} = \lambda_f \frac{\partial f}{\partial h};$$

$$\lambda_g = \frac{\partial L}{\partial g} = \lambda_h \frac{\partial h}{\partial g}$$

$$x \to g \to h \to f \to L$$

$$\frac{\partial L}{\partial x} = \lambda_g \frac{\partial g}{\partial x} \leftarrow \lambda_g = \frac{\partial L}{\partial g} = \lambda_h \frac{\partial h}{\partial g} \leftarrow \lambda_h = \frac{\partial L}{\partial h} = \lambda_f \frac{\partial f}{\partial h} \leftarrow \lambda_f = \frac{\partial L}{\partial f}$$



$$p(\alpha, \underline{x}, z) = P(z)p(\underline{x} \mid z)P(\alpha \mid \underline{x})$$

$$ECM = \sum_{j=1}^{C}\sum_{i=0}^{C} \lambda_{ij} P(\alpha = i, z = j)$$

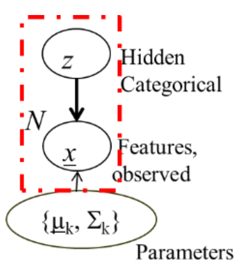
$$= \int_{\underline{x}} \sum_{i=0}^{C} P(\alpha = i \mid \underline{x}).\left[\sum_{j=1}^{C} \lambda_{ij}\, p(\underline{x} \mid z = j).P(z = j)\right] d\underline{x}$$

$$\Rightarrow \text{ Pick action } \alpha = k \text{ (class } \hat{z} = k\text{), if } k = arg \min_{i \in \{0,1,2,\ldots,C\}} \sum_{j=1}^{C} \lambda_{ij}\, p(\underline{x} \mid z = j)P(z = j)$$

$$= arg \min_{i \in \{0,1,2,\ldots,C\}} \sum_{j=1}^{C} \lambda_{ij}\, p(z = j \mid \underline{x})$$

Posterior probabilities or densities of unknowns are sufficient statistics for decision-making.

Problem: Do not know $\{P(z = j), p(\underline{x} \mid z = j), \lambda_{ij}\}$



$$QDA: \max g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x} - \underline{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(z = i)$$

$$LDA: g_i(\underline{x}) = \underline{\mu}_i^T \Sigma^{-1}\underline{x} - [\frac{1}{2}\underline{\mu}_i^T \Sigma^{-1}\underline{\mu}_i - \ln P(z = i)]$$

$$= \underline{w}_i^T \underline{x} - w_{i0} \quad \text{....linear rule}$$

sigmoid, *soft max* for posteriors

$$P(z = i \mid \underline{x}) \propto \exp(\alpha g_i(\underline{x}))$$

Discuss ML, Bayesian and discriminant approaches: Draw the pictures.

$ML : \theta = \{ \underline{\mu}_k, \Sigma_k \}$ are constant, but unknown parameters;

$\underline{\theta} = \{ (\underline{\mu}_k, \Sigma_k) \}$ General Case or $(\{ \underline{\mu}_k \}, \Sigma)$ Hyperellipsoid Case or $(\{ \underline{\mu}_k \}, \sigma^2 I_p)$ Hypersphere case

Bayesian: $\{ \underline{\mu}_k, \Sigma_k \}$ are random with known distributions

e.g., $p(\underline{\mu}_k | \Sigma_k) \sim N(\underline{m}_0, \Sigma_k / k_0); \; p(\Sigma_k) \sim IW(\Sigma_{0k}, \nu_{0k})$...Generalization of gamma and chi-squared

$$IW(\Sigma_v | \Sigma_0, \nu_0) = \frac{|\Sigma_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}} \Gamma_p(\frac{\nu_0}{2})} |\Sigma_v|^{-\frac{\nu_0 + p + 1}{2}} e^{-\frac{1}{2} tr(\Sigma^0 \Sigma_v^{-1})}; \; \Gamma_p(\frac{\nu_0}{2}) = \prod_{i=1}^{p} \Gamma(\frac{\nu_0 + 1 - i}{2});$$

Why learn the parameters or why generative learning?

Why not learn the weights in QDA and LDA or posterior probabilities directly? Discriminative learning.

$$\{P(z = k)\}_{k=1}^{C} \qquad \{ p(\underline{x}, \underline{\theta} | z = k) \}_{k=1}^{C}$$

$$D = \{ \underline{x}_k^1 \; \underline{x}_k^2 \; \underline{x}_k^3 .... \underline{x}_k^{n_k} : \; k = 1, 2, 3, \; .... C \} \quad n_k \text{ samples from class } k. \quad \text{Let } \sum_{k=1}^{C} n_k = N$$

$$L(\underline{\theta}) = p(D|\underline{\theta}) = \prod_{k=1}^{C} \prod_{j=1}^{n_k} p(\underline{x}_k^j | z = k, \underline{\theta}) P(z = k)$$

$$l(\underline{\theta}) = \ln L(\underline{\theta}) = \ln p(D | \underline{\theta})$$

$$= \sum_{k=1}^{C} \sum_{j=1}^{n_k} \ln p(\underline{x}_k^j | z = k, \underline{\theta}) + \sum_{k=1}^{c} n_k \ln \pi_k ; P(z = k) = \pi_k$$

$$\max \sum_{k=1}^{C} n_k \ln \pi_k \qquad \text{s.t.} \quad \sum_{k=1}^{C} \pi_k = 1$$

$$Lagrangian : \max_{\{\pi_k\}} [\sum_{k=1}^{C} n_k \ln \pi_k + \lambda (\sum_{k=1}^{C} \pi_k - 1)]$$

$$\frac{n_k}{\hat{\pi}_k} = -\lambda \quad \Rightarrow \quad \hat{\pi}_k = -\frac{n_k}{\lambda} \quad \Rightarrow \quad \lambda = -N$$

$$\boxed{\hat{\pi}_k = \frac{n_k}{N}}$$

Laplacian smoothing to address black swan problem: $\hat{\pi}_k = \dfrac{n_k + 1}{N + C}$

Bayesian interpretation:

$$L(\{\pi_k\}) = \prod_{k=1}^{C} \pi_k^{n_k}$$

$$p(\underline{\pi} | \underline{\alpha}) = Dir(\underline{\pi} | \underline{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)....\Gamma(\alpha_C)} \prod_{k=1}^{C} \pi_k^{\alpha_k - 1}; \alpha_0 = \sum_{k=1}^{C} \alpha_k$$

$$\boxed{\hat{\pi}_k^{MMSE} = \frac{n_k + \alpha_k}{N + \alpha_0}}$$

$$p(\underline{\pi} | D^N, \underline{\alpha}) = \frac{p(D^N | \underline{\pi}).p(\underline{\pi} | \underline{\alpha})}{p(D^N | \underline{\alpha})} = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + n_1).....\Gamma(\alpha_C + n_C)} \prod_{k=1}^{C} \pi_k^{\alpha_k + n_k - 1}$$

$$= \text{Dir}(\underline{\pi} | \underline{\alpha} + \underline{n})$$

$$\boxed{\hat{\pi}_k^{MAP} = \frac{n_k + \alpha_k - 1}{N + \alpha_0 - C}}$$

$$l(\{\underline{\mu}_k\}_{k=1}^{C}, \Sigma) = -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{k=1}^{C} \sum_{j=1}^{n_k} (\underline{x}_k^j - \underline{\mu}_k)^T \Sigma^{-1} (\underline{x}_k^j - \underline{\mu}_k)$$

$$\nabla_{\underline{\mu}_k} l = \underline{0} \quad \Rightarrow \quad \sum_{j=1}^{n_k} \hat{\Sigma}^{-1} (\underline{x}_k^j - \hat{\underline{\mu}}_k) = 0 \Rightarrow \hat{\underline{\mu}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \underline{x}_k^j ; k = 1, 2, .., C$$

$$\nabla_{\Sigma} l = \underline{0}$$

$$\nabla_{\Sigma}\left[\ln|\Sigma|\right] = \Sigma^{-1}$$

$$\nabla_{\Sigma} l = -\frac{1}{2}N\hat{\Sigma}^{-1} + \frac{1}{2}\hat{\Sigma}^{-1}\left[\sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\hat{\mu}}_k)(\underline{x}_k^j - \underline{\hat{\mu}}_k)^T\right]\hat{\Sigma}^{-1} = \underline{0}$$

$$\boxed{\begin{aligned} &Aside: A \text{ is } PD \\ &|A| = \exp(\ln|A|) = \exp\left(\sum_{i=1}^{n}\ln\lambda_i\right) = \exp(trace[\ln A]) \\ &\Rightarrow \ln|A| = trace[\ln A] \\ &\nabla_A\left[\ln|A|\right] = A^{-1} \end{aligned}}$$

$$\boxed{\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\hat{\mu}}_k)(\underline{x}_k^j - \underline{\hat{\mu}}_k)^T}$$

$$(\underline{x}_k^j - \underline{\hat{\mu}}_k)(\underline{x}_k^j - \underline{\hat{\mu}}_k)^T$$

$$\boxed{E\left[\hat{\Sigma}\right] = \frac{N-C}{N}\Sigma}$$

$$\boxed{\hat{\Sigma} = \frac{1}{N-C}\sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\hat{\mu}}_k)(\underline{x}_k^j - \underline{\hat{\mu}}_k)^T}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{n_k}(\underline{x}_k^j - \underline{\hat{\mu}}_k)(\underline{x}_k^j - \underline{\hat{\mu}}_k)^T \quad know\ n_k\underline{\hat{\mu}}_k = \sum_{j=1}^{n_k}\underline{x}_k^j$$

$$= \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{n_k}[\underline{x}_k^j\underline{x}_k^{jT} - \underline{x}_k^j\underline{\hat{\mu}}_k{}^T - \underline{\hat{\mu}}_k\underline{x}_k^{jT} + \underline{\hat{\mu}}_k\underline{\hat{\mu}}_k{}^T]$$

$$= \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{n_k}[\underline{x}_k^j\underline{x}_k^{jT} - \underline{\hat{\mu}}_k\underline{\hat{\mu}}_k{}^T]$$

$$E(\hat{\Sigma}) = \Sigma + \frac{1}{N}\sum_{k=1}^{C}n_k\underline{\mu}_k\underline{\mu}_k{}^T - \frac{1}{N}E\{\sum_{k=1}^{C}\frac{n_k}{n_k^2}\sum_{q=1}^{n_k}\sum_{r=1}^{n_k}\underline{x}_k^q\underline{x}_k^{rT}\}$$

$$= \Sigma + \frac{1}{N}\sum_{k=1}^{C}n_k\underline{\mu}_k\underline{\mu}_k{}^T - \frac{1}{N}\sum_{k=1}^{C}n_k\underline{\mu}_k\underline{\mu}_k{}^T - \frac{C}{N}\Sigma$$

$$= \left(\frac{N-C}{N}\right)\Sigma$$

$$\boxed{\hat{\Sigma}_k(\alpha) = \frac{(1-\alpha)n_k\hat{\Sigma}_k + \alpha N\hat{\Sigma}}{(1-\alpha)n_k + \alpha N}}$$

$$\boxed{\hat{\Sigma}(\gamma) = (1-\gamma)\hat{\Sigma} + \gamma I; \quad 0 < \gamma < 1}$$

$$\boxed{\hat{\Sigma}(\gamma) = \gamma\, diag\left(\hat{\Sigma}\right) + (1-\gamma)\hat{\Sigma}}$$

$$\boxed{\hat{\Sigma}_k(\alpha,\gamma) = (1-\gamma)\hat{\Sigma}_k(\alpha) + \frac{\gamma}{p}tr(\hat{\Sigma}_k(\alpha))I}$$

SA: Useful for streaming non-stationary data

$$\underline{\hat{\mu}}_k^n = \underline{\hat{\mu}}_k^{n-1} + \alpha_n(\underline{x}^n - \underline{\hat{\mu}}_k^{n-1})$$

$$\Sigma_k^n = \Sigma_k^{n-1} + \alpha_n[(\underline{x}^n - \underline{\hat{\mu}}_k^{n-1})(\underline{x}^n - \underline{\hat{\mu}}_k^{n-1})^T - \Sigma_k^{n-1}]$$

SA Conditions:

$$\lim_{n\to\infty}\alpha_n = 0$$

$$\sum_{n=1}^{\infty}\alpha_n = \infty \qquad \sum_{n=1}^{\infty}\alpha_n{}^2 < \infty$$

Bayesian:

$$P(z=k\,|\,\underline{x}, D_k) = \frac{p(\underline{x}\,|\,z=k, D_k)P(z=k)}{\sum_{i=1}^{C}p(\underline{x}\,|\,z=i, D_i)P(z=i)}$$

$$p(\underline{x}\,|\,z=k, D_k) = \int_{\underline{\theta}} p(\underline{x}\,|\,\underline{\theta}, z=k)\,p(\underline{\theta}\,|\,z=k, D_k)\,d\underline{\theta}$$

Tough to compute unless reproducing density. Need Simulation.

$$p(\underline{\theta}\,|\,z=k, D_k) = \frac{p(D_k\,|\,z=k,\underline{\theta})\,p(\underline{\theta}\,|\,z=k)}{\int_{\underline{\theta}} p(D_k\,|\,z=k,\underline{\theta})\,p(\underline{\theta}\,|\,z=k)\,d\underline{\theta}}$$

$$= \frac{\prod_{j=1}^{n_k}p(\underline{x}_k^j\,|\,z=k,\underline{\theta})\,p(\underline{\theta}\,|\,z=k)}{\int_{\underline{\theta}}\prod_{j=1}^{n_k}p(\underline{x}_k^j\,|\,z=k,\underline{\theta})\,p(\underline{\theta}\,|\,z=k)\,d\underline{\theta}}$$

If $p(D_k\,|\,z=k,\underline{\theta})$ has sharp peak at
then so does $p(\underline{\theta}\,|z=k, D_k)$. May be MAP

$$Let\ D_k^{n_k} = \{\underline{x}_k^1, \underline{x}_k^2, \ldots\ldots, \underline{x}_k^{n_k-1}, \underline{x}_k^{n_k}\} = \{D_k^{n_k-1}, \underline{x}_k^{n_k}\}$$

$$p(D_k^{n_k} \mid z = k, \underline{\theta}) = p(\underline{x}_k^{n_k} \mid z = k, \underline{\theta})\, p(D_k^{n_k-1} \mid z = k, \underline{\theta})$$

$$p(\underline{\theta} \mid z = k, D_k^{n_k}) = \frac{p(\underline{x}_k^{n_k} \mid z = k, \underline{\theta})\, p(\underline{\theta} \mid z = k, D_k^{n_k-1})}{\int p(\underline{x}_k^{n_k} \mid z = k, \underline{\theta})\, p(\underline{\theta} \mid z = k, D_k^{n_k-1})\, d\underline{\theta}}$$

$$where\ p(\underline{\theta} \mid z = k, D_k^0) = p(\underline{\theta} \mid z = k)$$

For exponential family (e.g., Gaussian, exponential, Rayleigh, Gamma, Beta, Poisson, Bernoulli, Binomial, Multinomial) need only few parameters to characterize the density. They are called *sufficient statistics*.

Application to Gaussian case:

$$p(\underline{x}^n \mid \underline{\mu}, \Sigma_v) = N(\underline{\mu}, \Sigma_v)\ and\ p(\underline{\mu}, \Sigma_v) = \overbrace{\underbrace{p(\underline{\mu} \mid \underline{m}_0, \frac{1}{k_0}\Sigma_v)}_{Gaussian}\ \underbrace{p(\Sigma_v \mid \Sigma_0, \nu_0)}_{inverse-Wishart}}^{NIW\,(Normal-Inverse-Wishart)}$$

$$Given\ D^n = \{\underline{x}^1, \underline{x}^2, \ldots., \underline{x}^n\},\ p(\underline{\mu}, \Sigma_v \mid D^n) = NIW(\underline{\mu}, \Sigma_v \mid \underline{m}_n, k_n, \nu_n, \Sigma_n)$$

$$where\ \underline{m}_n = \frac{k_0}{k_0 + n}\underline{m}_0 + \frac{n}{k_0 + n}\overline{\underline{x}}^n = \frac{k_{n-1}}{k_n}\underline{m}_{n-1} + \frac{1}{k^n}\underline{x}^n$$

$$k_n = k_0 + n = k_{n-1} + 1;\ \nu_n = \nu_0 + n = \nu_{n-1} + 1$$

$$\Sigma_n = \Sigma_0 + \sum_{i=1}^{n}(\underline{x}^i - \overline{\underline{x}}^n)(\underline{x}^i - \overline{\underline{x}}^n)^T + \frac{\nu_0 n}{\nu_n}(\overline{\underline{x}}^n - \underline{m}_0)(\overline{\underline{x}}^n - \underline{m}_0)^T$$

Density estimators

Relate to KNN

PNN

Voronoi Diagrams

Delaunay triangles

Sigmoid, Logistic regression, Softmax

IRLS

Discuss Laplace approximation

Probit approximation to posterior