$$p(\underline{x},\underline{y}) = N(\begin{bmatrix} \underline{\mu}_x \\ \underline{\mu}_y \end{bmatrix}; \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix})$$

$$\Rightarrow p(\underline{x}) = N(\underline{\mu}_x, \Sigma_{xx}); p(\underline{y}) = N(\underline{\mu}_y, \Sigma_{yy})$$

$$p(\underline{x} \mid \underline{y}) = N(\underline{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\underline{y} - \underline{\mu}_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X) = E_{p(\underline{x})}[-\ln p(\underline{x})] = \frac{1}{2}[n_x \ln(2\pi e) + \ln|\Sigma_{xx}|]; H(Y) = \frac{1}{2}[n_y \ln(2\pi e) + \ln|\Sigma_{yy}|]$$

$$H(X,Y) = \frac{1}{2}[(n_x + n_y)\ln(2\pi e) + \ln|\Sigma_{xx}| + \ln|\Sigma_{yy} - \Sigma_{xy}^T\Sigma_{xx}^{-1}\Sigma_{xy}|]$$

$$= \frac{1}{2}[(n_x + n_y)\ln(2\pi e) + \ln|\Sigma_{yy}| + \ln|\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T|]$$

$$I(X;Y) = \frac{1}{2}[\ln|\Sigma_{xx}| - \ln|\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T|] = \frac{1}{2}[\ln|\Sigma_{yy}| - \ln|\Sigma_{yy} - \Sigma_{xy}^T\Sigma_{xx}^{-1}\Sigma_{xy}|]$$

$$x, y\ scalars: \Sigma_{xx} = \sigma_x^2; \Sigma_{xy} = \rho_{xy}\sigma_x\sigma_y; \Sigma_{yy} = \sigma_y^2$$

$$I(X;Y) = \frac{1}{2}[\ln\sigma_x^2 - \ln(\sigma_x^2(1-\rho_{xy}^2))] = -\frac{1}{2}\ln(1-\rho_{xy}^2) = \ln\frac{1}{\sqrt{1-\rho_{xy}^2}}$$

Why sigmoid and softmax?

$$P(z=1 \mid \underline{x}) = \frac{p(\underline{x} \mid z=1)P(z=1)}{p(\underline{x})} = \frac{p(\underline{x} \mid z=1)P(z=1)}{p(\underline{x} \mid z=1)P(z=1) + p(\underline{x} \mid z=0)P(z=0)}$$

$$= \frac{1}{1+\exp(-\ln\frac{p(\underline{x} \mid z=1)}{p(\underline{x} \mid z=0)} - \ln\frac{P(z=1)}{P(z=0)})} = \frac{1}{1+\exp(-h(x,\underline{w}))}$$

$$\Rightarrow P(z=1 \mid \underline{x}) = \frac{\exp(f(x,w_1))}{\exp(f(x,w_1)) + \exp(f(x,w_0))} = \frac{1}{1+\exp(-[(f(x,w_1) - f(x,w_0)]}$$

$$P(z=i \mid \underline{x}) = \frac{\exp(f(x,w_i))}{\sum_{j=}^{C}\exp(f(x,w_j))} = s_i(x,\underline{w}) ... soft\ \max$$

$$\nabla_{w_i} s_i(x,\underline{w}) = s_i(x,\underline{w})(1 - s_i(x,\underline{w})); \nabla_{w_j} s_i(x,\underline{w}) = -s_i(x,\underline{w})s_j(x,\underline{w})); j \neq i$$

$$-\ln P(z_n \mid \underline{x}_n, \underline{w}) = -z_n \ln \underbrace{P(z_n = 1 \mid \underline{x}_n, \underline{w})}_{g(y_n)} - (1-z_n)\ln \underbrace{P(z_n = 0 \mid \underline{x}_n, \underline{w})}_{1-g(y_n)}$$

$$= -z_n \ln g(y_n) - (1-z_n)\ln(1-g(y_n))$$

Data Distribution: $p = z_n \in \{0,1\}$; Classifier output: $q = \{g(y_n), 1 - g(y_n)\}$

Cross-entropy $H_n(p,q) = -z_n \ln g(y_n) - (1 - z_n)\ln(1 - g(y_n))$

$$= -z_n \ln z_n - (1 - z_n)\ln(1 - z_n)$$

$$+ [z_n \ln \frac{z_n}{g(y_n)} + (1 - z_n)\ln \frac{(1 - z_n)}{(1 - g(y_n))}]$$

$$= H_n(p) + KL(p \| q)$$

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

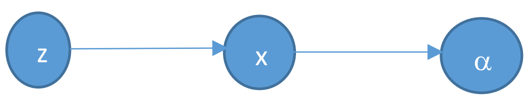
$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial x} = \lambda_g \frac{\partial g}{\partial x}$$

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

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

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

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



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

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

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

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

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

$\Lambda = \begin{bmatrix}\lambda_{01} & \lambda_{02}\\ \lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{bmatrix};\lambda_{11} < \lambda_{12};\lambda_{22} < \lambda_{21} \Rightarrow$ correct decisions have less cost

Since $P(z = 2\mid\underline{x}) = 1 - P(z = 1\mid\underline{x})$, the decision rule is:

$k = arg \min_{i\in\{0,1,2\}}\{(\lambda_{i1} - \lambda_{i2})P(z = 1\mid\underline{x}) + \lambda_{i2}\}$

Special case 1: $\lambda_{01} = \lambda_{02} = \lambda_r = \infty \Rightarrow$ Likelihood Ratio Rule

$\dfrac{p(\underline{x}\mid z = 1)}{p(\underline{x}\mid z = 2)} \geq \dfrac{(\lambda_{12} - \lambda_{22})P(z = 2)}{(\lambda_{21} - \lambda_{11})P(z = 1)} \Rightarrow \hat{z} = 1;$ otherwise $\hat{z} = 2$ (Prove it)

Special case 2: $\lambda_{01} = \lambda_{02} = \lambda_r$

Reject range for $P(z = 1\mid\underline{x})$ exists if $\dfrac{\lambda_r - \lambda_{12}}{\lambda_{11} - \lambda_{12}} > \dfrac{\lambda_r - \lambda_{22}}{\lambda_{21} - \lambda_{22}}$; else no reject decision (Prove it)

Special case 3: $\lambda_{01} = \lambda_{02} = \lambda_r;\lambda_{11} = \lambda_{22} = 0;\lambda_{12} > 0;\lambda_{21} > 0$

Reject range for $P(z = 1\mid\underline{x})$ exists if $\dfrac{1}{\lambda_r} > \dfrac{1}{\lambda_{12}} + \dfrac{1}{\lambda_{21}}$; else no reject decision (Prove it)

Pick action $\alpha = k, k \in \{0,1,2,..,C\}$

If $k = \arg\min\{\lambda_r, \lambda_e[1 - P(z = 1\mid\underline{x})], ..., \lambda_e[1 - P(z = C\mid\underline{x})]\}$

$$= \arg\min\{\frac{\lambda_r}{\lambda_e}, [1 - P(z = 1\mid\underline{x})], ..., [1 - P(z = C\mid\underline{x})]\} \because \lambda_e > 0$$

$$= \arg\min\{\frac{\lambda_r}{\lambda_e} - 1, -P(z = 1\mid\underline{x}), ..., -P(z = C\mid\underline{x})\}$$

$$= \arg\max\{\underbrace{1 - \frac{\lambda_r}{\lambda_e}}_{\beta}, P(z = 1\mid\underline{x}), ..., P(z = C\mid x)\}$$

Decide for $\alpha = k$ (*class* $\hat{z} = k$): $\begin{cases} if \;\; P(z = k\mid\underline{x}) = \max_{j\in\{1,2,..,C\}}P(z = j\mid\underline{x}) \;\underline{and}\; P(z = k\mid\underline{x}) > \beta \\ \qquad otherwise \qquad reject\,(action\; \alpha = 0) \end{cases}$

MAP: $\beta = 0 \Rightarrow \lambda_r = \lambda_e \Rightarrow arg \max_{j\in\{1,2,...,C\}}P(z = j\mid\underline{x})$

ML: If $p(z=j)=\frac{1}{C}\ \forall j \ \Rightarrow \arg \max_{j\in\{1,2,..,C\}} \ln p(\underline{x}\,|\,z=j)$

Properties of Gaussians: Sums, Linear Transformations, Conditionals

$$\underline{x}=\begin{bmatrix}\underline{x}_1\\ \underline{x}_2\end{bmatrix} \text{ and } \underline{x}\sim N(\underline{\mu},\Sigma);\ \underline{\mu}=\begin{bmatrix}\underline{\mu}_1\\ \underline{\mu}_2\end{bmatrix};\ \Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{bmatrix};\ J=\Sigma^{-1}=\begin{bmatrix}J_{11} & J_{12}\\ J_{12}^T & J_{22}\end{bmatrix}$$

$$J_{11}=(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1};\ J_{22}=(\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1};\ J_{12}=-(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

Then, the marginal & conditional densities are also Gaussian

$$p(\underline{x}_2)=N(\underline{\mu}_2,\Sigma_{22})$$

$$p(\underline{x}_1\,|\,\underline{x}_2)=N(\underline{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2-\underline{\mu}_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)$$

$$=\mathrm{N}(\underline{\mu}_1-J_{11}^{-1}J_{12}(\underline{x}_2-\underline{\mu}_2),J_{11}^{-1})$$

- Covariance matrix captures marginal independencies between variables

$$\Sigma_{ij}=0 \Leftrightarrow x_i \ \&\ x_j \ \textit{are independent}\ (or)\ x_i \perp x_j$$

- Information matrix $J=\Sigma^{-1}$ captures conditional independencies

  $$J_{ij}=0 \Leftrightarrow x_i \perp x_j \ \Big|\Big\{\{x_1,x_2,...,x_n\}-\{x_i,x_j\}\Big\}$$

  Non-zero entries in J correspond to edges in the dependency network
- Simulating multivariate Gaussian random variables
- Discriminants

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

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

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

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

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

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

  $\textit{sigmoid}, \textit{soft max}$ for posteriors

  $$g_i(\underline{x})=\frac{1}{\sigma^2}\underline{\mu}_i^T\underline{x}-(\frac{1}{2\sigma^2}\underline{\mu}_i^T\underline{\mu}_i-\ln P(z=i))$$

  $$= \quad \underline{w}_i^T\underline{x} \quad - \quad w_{io}$$

  $$g(\underline{x})=\sigma^2[g_1(\underline{x})-g_2(\underline{x})]$$

- $$=(\underline{\mu}_1-\underline{\mu}_2)^T\underline{x}-[\frac{1}{2}(\underline{\mu}_1^T\underline{\mu}_1-\underline{\mu}_2^T\underline{\mu}_2)-\sigma^2\ln\frac{P(z=1)}{P(z=2)}]$$

  $$=\underline{w}^T\underline{x}-w_o=\underline{w}^T(\underline{x}-\underline{x}_o)=0;\ \underline{x}_o=\frac{\underline{w}}{\underline{w}^T\underline{w}}w_0$$

Chernoff Bound:

$$P(error)=\int_{\underline{x}} P(error\,|\,\underline{x})p(\underline{x})d\underline{x}$$
$$=\int_{\underline{x}} min[P(z=1|\underline{x}),P(z=2|\underline{x})]\,p(\underline{x})d\underline{x}$$
$$\leq [P(z=1)]^{\beta}[P(z=2)]^{1-\beta}\int_{\underline{x}}[p(\underline{x}|z=1)]^{\beta}[p(\underline{x}|z=2)]^{1-\beta}d\underline{x}$$

$$If\ \ a \geq b\ then\ (a/b)^{\beta} \geq 1$$
$$\Rightarrow (a/b)^{\beta} b \geq b$$
$$\Rightarrow a^{\beta} b^{1-\beta} \geq b$$

$$P(error)\leq [P(z=1)]^{\beta}[P(z=2)]^{1-\beta}e^{-k(\beta)}=e^{-k(\beta)+\beta ln P(z=1)+(1-\beta)ln[1-P(z=1)]}$$

Gaussian:
$$where\quad k(\beta)=\frac{\beta(1-\beta)}{2}||\underline{\mu}_2-\underline{\mu}_1||^2_{[\beta\Sigma_1+(1-\beta)\Sigma_2]^{-1}}+\frac{1}{2}ln\frac{|\beta\Sigma_1+(1-\beta)\Sigma_2|}{|\Sigma_1|^{\beta}|\Sigma_2|^{1-\beta}}$$

$$P(error)\leq\sqrt{P(z=1)P(z=2)}\,e^{-k(1/2)}$$

Bhattacharyya:
$$where\quad k(1/2)=\frac{1}{8}||\underline{\mu}_2-\underline{\mu}_1||^2_{[\frac{\Sigma_1+\Sigma_2}{2}]^{-1}}+\frac{1}{2}ln\frac{|\frac{\Sigma_1+\Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

Discuss ROC

Binary case … write as graph

Missing features

Noisy features

- Generative versus Discriminative?
- Can we exploit & generalize the forms of discriminant functions? Can we estimate discriminants directly?
- Different types of learning (unsupervised, supervised, semi-supervised, RL)
- How to handle missing data?
- How do we select features x for best classification/regression accuracy?
- Can we exploit dependency structure among features?
- What happens if classes and features change dynamically?
- How do we validate data driven models? How do we select the best model?