GAUSSIAN MODELS

MVN: $N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma \cdot (x-\mu)\right]$

- EXPONENT: MAHALANOBIS DISTANCE BETWEEN DATA x AND MEAN VECTOR μ
- EIGENDECOMPOSITION: $\Sigma = U \Lambda U^T$

EIGEN VECTORS ARE 'ELLIPSE' DIRECTIONS

MLE: $\hat{\mu} = \bar{x}$ $\quad \hat{\Sigma} = \frac{1}{N}\left(\sum x_i x_i^T\right) - \bar{x}\bar{x}^T$ (EMPIRICAL MEAN AND COVARIANCE)

- HAS MAX ENTROPY AMONG DISTRIBUTIONS WITH SPECIFIED COVARIANCE Σ

GAUSSIAN DISCRIMINANT ANALYSIS

- DEFINES CLASS CONDITIONAL DENSITIES IN A GENERATIVE CLASSIFIER $p(x|y=c,\theta) = N(x|\mu_c,\Sigma_c)$ ■ CCD ARE GAUSSIAN
- Σ DIAGONAL $\longrightarrow$ GAUSSIAN NAIVE BAYES (FEATS ARE CI)
- $\hat{y}(x) = \underset{c}{\text{ARGMAX}}\left[\log p(y=c|\pi) + \log p(x|\theta_c)\right]$
- MEASURES DISTANCE OF x FROM CENTER OF EACH CLASS μ_c, MAHALANOBIS DISTANCE $\longrightarrow$ 'NEAREST CENTROIDS CLASSIFIER'
- NORMALLY QUADRATIC, LINEAR WHEN Σ_c ARE TIED OR SHARED; $\Sigma_c = \Sigma$
- INDUCED LABELS HAVE SAME FORM AS LOGISTIC REGRESSION
- REGULARIZED LDA $\Sigma = \lambda\, \text{DIAG}(\Sigma_{MLE}) + (1-\lambda)\Sigma_{MLE}$
- DIAGONAL LDA $\Sigma_c = \Sigma$, THEN DIAGONAL COV. MATRIX = TO RDA, $\lambda = 1$

POOL ACROSS CLASSES WORKS BETTER IN HIGH DIMENSIONS

- NEAREST SHRUNKEN CENTROIDS: MAP FOR DIAGONAL LDA WITH SPARCITY INDUCING PRIOR (LAPLACE)

CLASS SPECIFIC FEATURE MEAN $\mu_{cj} = m_j + \Delta_{cj}$

- RELATION TO LOGISTIC REGRESSION

MVN MARGINALS:

$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \quad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}$$

$$\begin{cases} p(x_1) = N(x_1|\mu_1, \Sigma_{11}) \\ p(x_2) = N(x_2|\mu_2, \Sigma_{12}) \end{cases}$$

MVN CONDITIONALS: $p(x_1|x_2) = N(x_1|\mu_{1|2}, \Sigma_{1|2})$

$$\mu_{1|2} = \Sigma_{1|2}\left(\Lambda_{11}\mu_1 - \Lambda_{12}(x_2-\mu_2)\right)$$

$$\Sigma_{1|2} = \Lambda_{11}^{-1}$$

2D MVN $\quad p(x_1|x_2) = N\left(x_1 \middle| \mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2-\mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$

GAUSSIAN INTERPOLATION:

D SUBINTERVALS

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + \epsilon_j, \quad \epsilon \sim N(0, 1/\lambda \cdot I)$$

$Lx = \epsilon$: , L IS FINITE DIFFERENCES MATRIX

- USED FOR DATA IMPUTATION

INFORMATION FORM OF MVN:

$\Lambda = \Sigma^{-1}$ & $\xi = \Sigma^{-1} \mu$

$$N_c(x|\xi, \Lambda) = (2\pi)^{-D/2} |\Lambda|^{1/2} \exp\left[-\frac{1}{2}\left(x^T \Lambda x + \xi^T \Lambda^{-1} \xi - 2x^T \xi\right)\right]$$

- Λ, ξ CANONICAL PARAMETERS
- μ, σ MOMENT PARAMETERS

- CONDITIONING IS EASIER! $P(x_1|x_2) = N_c(x_1|\xi_1 - \Lambda_{12}x_2 \Lambda_{11})$
- MULTIPLYING IS EASIER: $N_c(\xi_1, \Lambda_1) \cdot N_c(\xi_2, \Lambda_2) = N_c(\xi_1 + \xi_2, \Lambda_1 + \Lambda_2)$
- MARGINALIZATION EASIER IN MOMENT FORM

LINEAR GAUSSIAN SYSTEMS

PRIOR

$P(x) = N(x|\mu_x, \Sigma_x)$

$P(y|x) = N(y|Ax + b, \Sigma_y)$

~~POSTERIOR~~ LIKELIHOOD

~~LIKELIHOOD~~ POSTERIOR

$\longrightarrow P(x|y) = N(x|\mu_{x|y}, \Sigma_{x|y})$

$$\Sigma_{x|y}^{-1} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A, \quad \mu_{x|y} = \Sigma_{x|y}\left[A^T \Sigma_y^{-1}(y-b) + \Sigma_x^{-1} \mu_x\right]$$

$\longrightarrow P(y) = N(y|A\mu_x + b, \Sigma_y + A\Sigma_x A^T)$

DATA / NORMALIZATION

- USEFUL TO INFER SCALARS AND VECTORS FROM NOISY DATA, INTERPOLATE DATA
- TIKHONOV REGULARIZATION
- WISHART DISTRIBUTION

WISHART DISTRIBUTION

USED TO MODEL UNCERTAINTY IN COVARIANCE MATRICES Σ OR Λ, GENERALIZES Γ TO POSITIVE DEFINITE MATRICES, OR MULTIDIM GEN. OF χ^2

$$Wi(\Lambda|S, \nu) = \frac{1}{Z_{Wi}} |\Lambda|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} \mathrm{TR}(\Lambda S^{-1})\right)$$

ν = DOF

S = SCALE MATRIX

- ARE CONJUGATE PRIOR OF Λ
- Wi IS DISTRIB OF SAMPLE COV. MATRIX OF MVN

Z_{Wi} = NORMALIZATION $= 2^{\nu D/2} \Gamma_D(\nu/2) |S|^{\nu/2}$

$\Gamma_D(a)$ = MULTIVARIATE GAMMA

- ARE DISTRIBUTION OVER MATRICES

- $x_i \sim N(0, \Sigma) \rightarrow S = \sum_i^N x_i \cdot x_i^T$ IS WISHART $S \sim Wi(\Sigma, 1)$, $E[S] = N\Sigma$, MEAN $= \nu S$
- $D = 1 \rightarrow Wi(\lambda|s^{-1}, \nu) = Ga(\lambda|\frac{\nu}{2}, \frac{s}{2})$

INVERSE WISHART

- $IW(\Sigma|S, \nu) = \frac{1}{Z_{IW}} |\Sigma|^{-(\nu + D + 1)} \exp\left(-\frac{1}{2} \mathrm{TR}(S^{-1} \Sigma^{-1})\right)$ $\quad Z_{IW} = |S|^{-\nu/2} 2^{\nu D/2} \Gamma_D(\nu/2)$

- $D = 1$ $\quad IW(\sigma^2|s^{-1}, \nu) = IG(\sigma^2|\nu/2, s/2)$ $\quad$ INVERSE WISHART $\longrightarrow$ INVERSE GAMMA
- CONJ PRIOR OF Σ OF MVN

INFERRING MVN PARAMETERS

$x_i \sim \mathcal{N}(\mu, \Sigma)$

- POSTERIOR FOR μ

LIKELIHOOD $p(D|\mu) = \mathcal{N}(\bar{x} | \mu, \frac{1}{N}\Sigma)$

GAUSSIAN PRIOR $p(\mu) = \mathcal{N}(\mu | m_0, V_0)$

POSTERIOR (ALSO GAUSSIAN) $p(\mu | D, \Sigma) = \mathcal{N}(\mu | m_N, V_N)$

$$V_N^{-1} = V_0^{-1} + N\Sigma^{-1}$$

$$m_N = V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1} m_0)$$

IF UNINFORMATIVE PRIOR $\longrightarrow$ POSTERIOR MEAN = MLE

- POSTERIOR FOR Σ

$$p(\Sigma | D, \mu) = IW(\Sigma | S_N, \nu_N)$$

$\nu_N = \nu_0 + N$

$S_N^{-1} = S_0 + S_\mu$ SCATTER MATRIX

$$\hat{\Sigma}_{MAP} = \frac{S_0 + S_\mu}{N_0 + N}$$

REWRITE AS $\hat{\Sigma}_{MAP} = \lambda \Sigma_0 + (1-\lambda)\hat{\Sigma}_{MLE}$ FOR SHRINKAGE / REGULARIZED ESTIM.

BAYESIAN T-TEST

$$p(\mu|D) = T\left(\mu \middle| \bar{x}, \frac{s^2}{N}, N-1\right), \quad T = \frac{\bar{x}-\mu_0}{s/\sqrt{N}}, \quad P(\mu|D) = 1 - F_{N-1}(t)$$

- THERE μ IS UNKNOWN AND $\bar{x}$ IS FIXED, CONTRARY TO FREQUENTIST FRAMEWORK
- SAME FORM AS FREQUENTIST T-TEST IF UNINFORMATIVE PRIOR

FISHER'S LDA

PROJECT z IN LOWER DIM. SPACE $z = Wx$, $W = L \times D$

USUAL DA IS PROBLEMATIC IN HIGH DIMENSIONS. REDUCE DIMENSIONALITY. IE PCA, BUT SINCE IT'S UNSUPERVISED ITS RESULTS ARE NOT NECESSARILY OPTIMAL FOR CLASSIFICATION.

- **IDEA:** FIND W SO TO HAVE z BE OPTIMALLY CLASSIFIED WITH A GAUSSIAN CLASS CONDITIONAL DENSITY. GAUSSIANITY IS OK SINCE IT'S LINEAR COMBS OF STUFF

- **LDA** REDUCES TO AT MOST $L \leq C-1$ DIMENSIONS (BECAUSE RANK OF Σ_B IS $C-1$); IN TWO-CLASS CASE W IS A SINGLE VECTOR

- **2-CLASS CASE** $\mu_1 = \frac{1}{N_1} x_1$, $\mu_2 = \frac{1}{N_2} x_2$ $\quad m_k = W^T \mu_k$ PROJECTION OF MEAN ON LINE W. FIND W TO MAXIMIZE DISTANCE BETWEEN MEANS AND KEEP CLUSTERS TIGHT $J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$

  ... $S_B = \lambda S_W w \rightarrow$ GENERALIZED EIGENVALUE PROBLEM $\xrightarrow[\text{SHORTCUT}]{\text{2 CLASS}}$ $W = S_W^{-1}(\mu_2 - \mu_1)$

S_B = BETWEEN CLASS SCATTER, S_W = WITHIN-CLASS SCATTER

GENERAL: FIND EIGENVECTORS OF $S_W^{-1} S_B$

- **HIGHER DIMENSIONS, MOAR CLASSES** FIND MATRIX W TO MAXIMIZE $J(W) = \frac{W \Sigma_B W^T}{W \Sigma_W W^T}$ $\quad$ $W = \Sigma_W^{-1/2} U$. U IS L TOP EIGENVECTORS OF $\Sigma_W^{-1/2} \Sigma_B \Sigma_W^{-1/2}$

  $\Sigma_{B,W}$ = BETWEEN/WITHIN CLASS COV. MATRICES

- **EXTENSIONS:**
  - **HETEROSKEDASTIC LDA:** Σ_c ARE NOT DIAGONAL AND EQUAL = FLDA
  - **MULTIPLE LDA:** EACH CLASS HAS OWN PROJECTION MATRIX

- S_B: $\sum_i (\mu_i - \mu)(\mu_i - \mu)^T$
- S_W: $\sum_i \sum_j p_i (x_j - \mu_i)(x_j - \mu_i)^T$ = AUTOCOV $\times$ p_{class}

- **ALGO:** ~~CLASS~~ FIND FULL COV
  SELECT CLASS DATA
  FOREACH CLASS
  - FIND CLASS COV
  - SUM TO ~~COV~~ S_W

  $S_B = C - S_W$

  FIND EVECS

  PICK TOP EVECS $\rightarrow$ Ws

  PROJECT ORIGINAL DATA WITH W