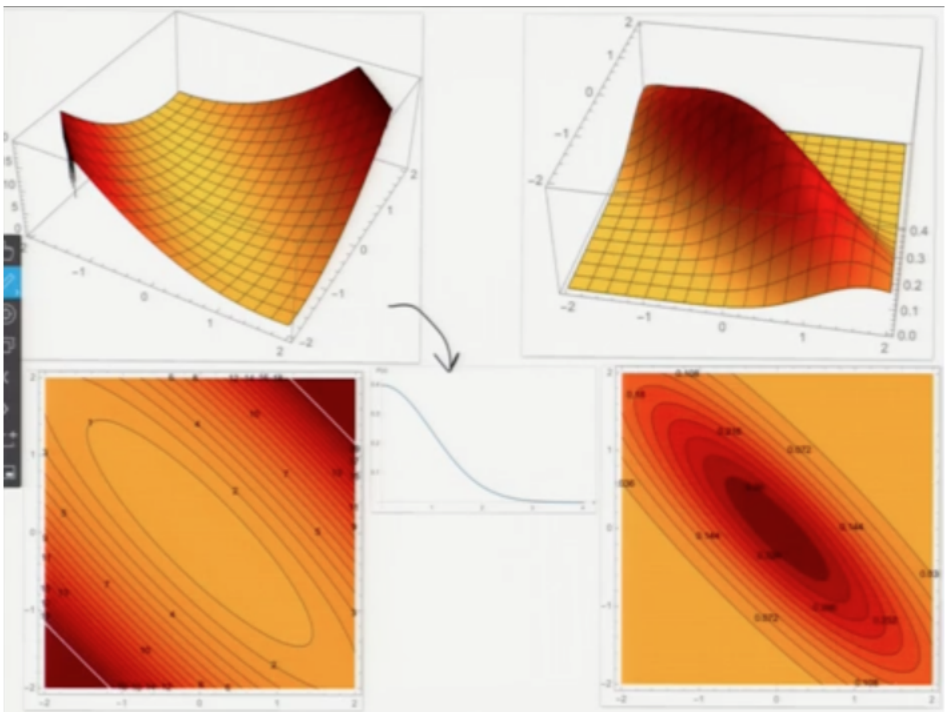
02/22/2016

Anisotropic Multivariate Gaussians

- $X \sim \mathcal{N}(\mu, \Sigma) \Leftarrow X$ is random d-vector with mean μ.

$$P(x) = \frac{1}{\sqrt{(2\pi)^d}\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

- Σ is the dxd SPD covariance matrix
- Σ^{-1} is the dxd SPD precision matrix; serves as a metric tensor.
- Write $P(x) = n(q(x))$, where $q(x) = (x-\mu)^T\Sigma^{-1}(x-\mu)$. Note, $n : \mathbb{R} \to \mathbb{R}$, exponential, $q : \mathbb{R}^d \to \mathbb{R}$, quadratic.
- Principle: given $f : \mathbb{R} \to \mathbb{R}$, isosurfaces of $f(q(x))$ are same as $q(x)$ (different isovalues), except that some might be "combined."



- Covariance:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])^T] \\ &= E[XY^T] - \mu_x\mu_y^T \\ \mathrm{Var}(X) &= \mathrm{Cov}(X,X) \end{aligned}$$

  - For a Gaussian, one can show $\mathrm{Var}(X) = \Sigma$. Hence,

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \dots & \mathrm{Cov}(X_1,X_d) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2,X_d) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_d,X_1) & \mathrm{Cov}(X_d,X_2) & \dots & \mathrm{Var}(X_d) \end{bmatrix}$$

- X_i, X_j independent $\Rightarrow \text{Cov}(X_i, X_j) = 0$.
- $\text{Cov}(X_i, X_j) = 0$ and they come from a joint normal distribution $\Rightarrow X_i, X_j$ independent.
- All features pairwise independent $\Rightarrow \Sigma$ is diagonal.

$$\Sigma \text{ is diagonal} \Leftrightarrow \text{axis-aligned Gaussian; squared radii on the diagonal.}$$
$$\Leftrightarrow P(x) = P(X_1)P(X_2)\dots P(X_d)$$

Covariance Matrix | Precision Matrix (Metric)

$$\begin{tikzcd}
\Sigma \arrow[rr, leftrightarrow, "\text{inverse}"] \arrow[d, shift right, "\sqrt{}"'] & & M = \Sigma^{-1} \arrow[d, shift right, "\sqrt{}"'] \\
\Sigma^{1/2} \arrow[rr, leftrightarrow, "\text{inverse}"] \arrow[u, shift right, "\text{square}"'] & & A = \Sigma^{-1/2} \arrow[u, shift right, "\text{square}"']
\end{tikzcd}$$

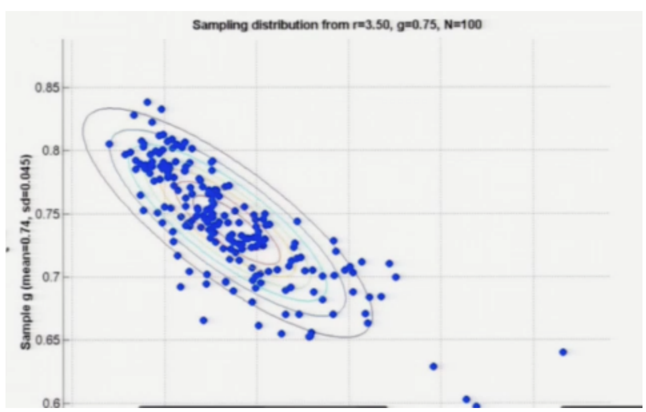
(Maps spheres to ellipsoids) | (Maps ellipsoids to spheres)

- Eigenvalues of $\Sigma^{1/2}$ are ellipsoid radii (standard deviations along the eigenvectors).
- Eigenvalues of Σ are variances along along eigenvectors.
- Diagonalizing $\Sigma = V\Lambda V^T$, $\Sigma^{\frac{1}{2}} = V\Lambda^{\frac{1}{2}}V^T$.

Maximum Likelihood estimation for anisotropic Gaussians

- Given samples $x_1, \dots, x_n$ and classes $y_1, \dots y_n$ find the best-fit Gaussians.
- For QDA:

$$\hat{\Sigma_c} = \frac{1}{n_c} \sum_{i:y_i=c} (x_i - \mu_c)(x_i - \mu_c)^T \Leftarrow \text{conditional covariance for samples in class c}$$



  - Priors π_c and means $\hat{\mu_c}$ same as before.
  - $\hat{\Sigma_c}$ is the positive semidefinite. If some zero eigenvalue, must eliminate the zero-variance dimension.

- For LDA:

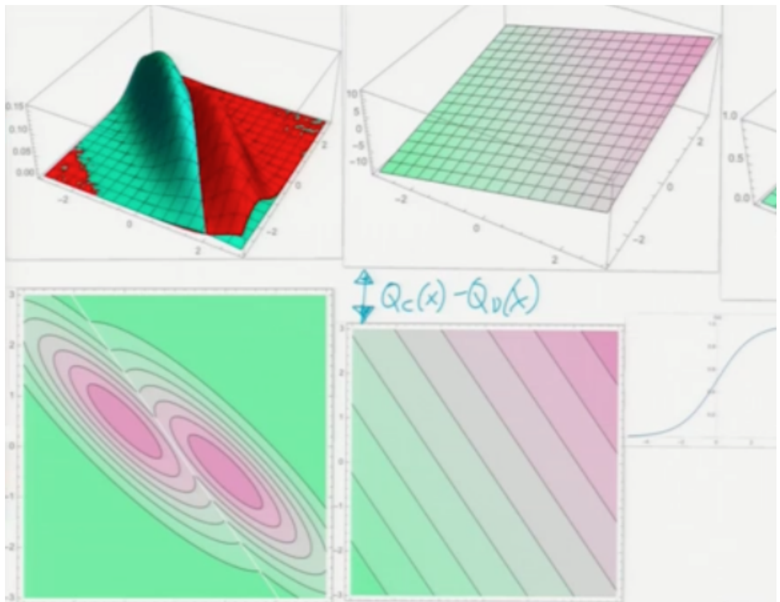
$$\hat{\Sigma} = \frac{1}{n} \sum_c \sum_{i:y_i=c} (x_i - \mu_c)(x_i - \mu_c)^T \Leftarrow \text{pooled within class covariance matrix}$$

- <u>QDA</u>:

- π_c, μ_c, Σ_c may be different for each class c.
- Goal is to choose c that maximizes $P(X = x|Y = c)\pi_c$, which is equivalent to maximizing the quadratic discriminant fu

$$
\begin{aligned}
Q_c(x) &= \ln\left(\sqrt{(2\pi)^d}P(x)\pi_c\right) \\
&= -\frac{1}{2}q_c(x) - \frac{1}{2}\ln|\Sigma_c| + \ln\pi_c
\end{aligned}
$$

- 2 classes: Prediction function $Q_c(x) - Q_d(x)$ is quadratic, but may be indefinite.
- Since the prediction function is quadratic $\Rightarrow$ Bayes decision boundary is quadric.
- Posterior is $P(Y = c|X = x) = s(Q_c(x) - Q_d(x))$ where $s(\cdot)$ is the logistic function.



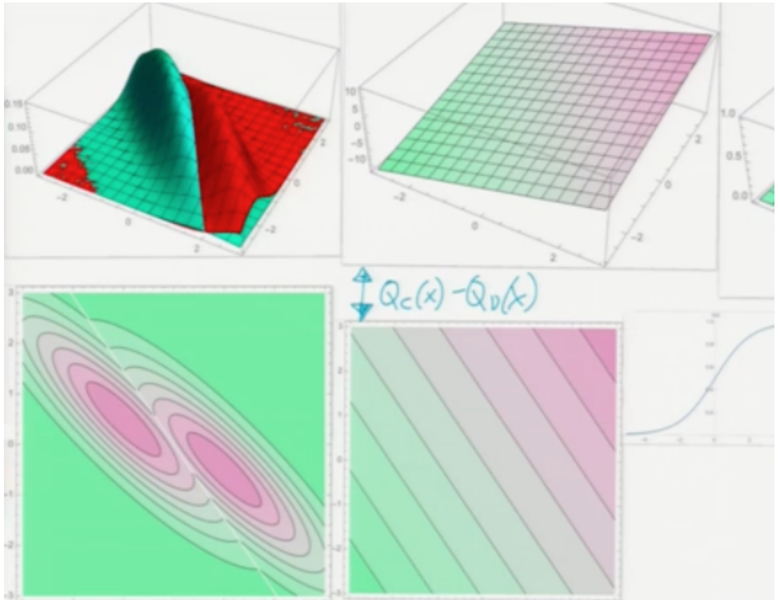
- LDA:
  - Once Σ for all classes.

$$
Q_c(x) - Q_d(x) = (\mu_c - \mu_d)^T\Sigma^{-1}x - \frac{\mu_c^T\Sigma^{-1}\mu_c - \mu_d^T\Sigma^{-1}\mu_d}{2} + \ln\pi_c - \ln\pi_d
$$

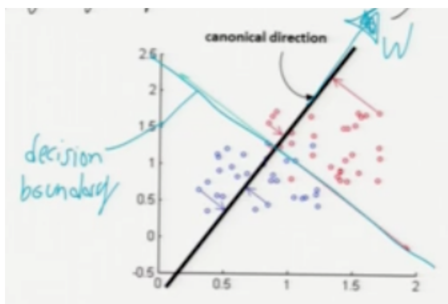
  - Choose class c that maximizes the linear discriminant function,

$$
\mu^T\Sigma^{-1}x - \frac{1}{2}\mu_c^T\Sigma^{-1}\mu_c + \ln\pi_c
$$

  - 2 classes:
    * Decision boundary is $w^Tx + \alpha = 0$
    * Posterior is $P(Y = c|X = x) = s(w^Tx + \alpha)$.



- <u>Notes</u>
  - Changing prior π_c (or loss) is easy: if its LDA adjust α.
  - LDA is often interpreted as projecting samples onto the normal vector.



  - For 2 classes,
    * LDA has $d + 1$ parameters (w, α).
    * QDA has $\frac{d(d+3)}{2} + 1$ parameters
    * $\Rightarrow$ QDA more likely to overfit.