Anisotropic Gaussians & GDA

from last lecture: PDF $f(x) = n(q(x)), n(q) = \frac{1}{\sqrt{2\pi d\,|\Sigma|}} e^{-q/2}$ ($\mathbb{R} \to \mathbb{R}$), $q(x) = (x-\mu)^T \Sigma^{-1} (x-\mu)$ ($\mathbb{R}^d \to \mathbb{R}$, quadratic)

↑ determinant of Σ

cov. matrix

eigendecomp: $\Sigma = V \Gamma V^T$

↑ evalues of Σ are variances σ_i^2

each variance associated w/ an eigenvector $\Gamma_{ii} = \sigma_i^2$

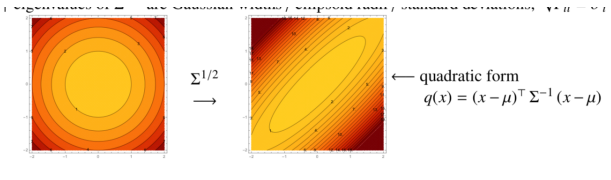
$\longrightarrow \Sigma^{1/2} = V \Gamma^{1/2} V^T$ maps spheres to ellipsoids

more helpful in terms of STDs

(widths of Gaussians in each i-th dimension)

↳ also radii of ellipsoids, $\sqrt{\Gamma_{ii}} = \sigma_i$

width of univariate gauss of std. $\sigma \to \sqrt{\gamma_i}$ of Σ



decomp Σ^{-1}: $\Sigma^{-1} = V \Gamma^{-1} V^T$ ← precision matrix / metric tensor

↳ defines isocontours of

Maximum Likelihood Estimation for Anistropic Gaussians

given $(x_1, y_1) \ldots (x_n, y_n)$, find best fit normal distr. for each class

*x_i is a column vector

sample mean

for QDA: $\hat{\Sigma}_c = \boxed{\frac{1}{n_c} \sum_{i; y_i = c}} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$

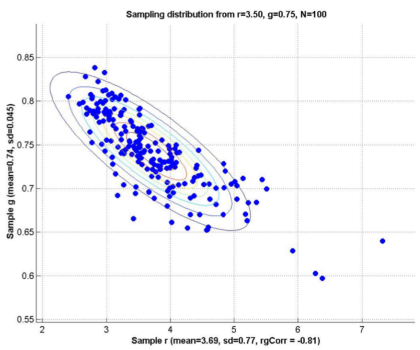
mean

rank 1 outer prod., $d \times d$

← conditional covariance matrix for points in class c

- prior $\hat{\pi}_c$ that a sample point is in class c
- mean $\hat{\mu}_c$ = avg of all points in c

Gaussian generated from points →



* recall $\hat{\Sigma}_c$ is PSD, but not always PD

What if I want to do LDA?

- In LDA, every class has same covariance

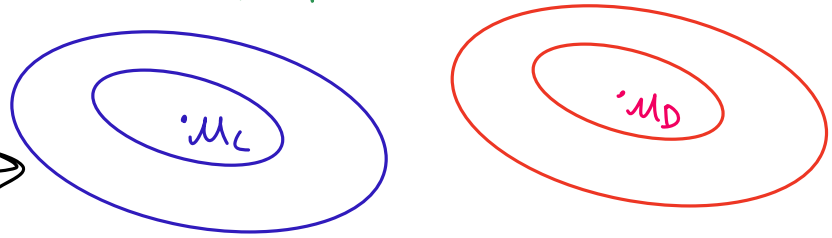
Within class $\Rightarrow$ using mean for each class

LDA: $\hat{\Sigma} = \frac{1}{n} \sum_c \sum_{i:y_i=c} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$ ← Pooled-within-class Covariance matrix

over all classes

contribution of each sample points within this class

Same variances, different means →



If not within-class, we'd get 1 big contour

QDA: choose class C that maximizes $f(X=x|Y=c)\pi_c$ ← $l(\cdot,\cdot)$

Same as maximizing the quadratic discriminant function

$$Q_c(x) = \ln\left((\sqrt{2\pi})^d f_c(x)\pi_c \right) = -\frac{1}{2}(x-\mu_c)^T \hat{\Sigma}_c^{-1}(x-\mu_c) - \frac{1}{2}\ln|\hat{\Sigma}_c| + \ln\pi_c$$

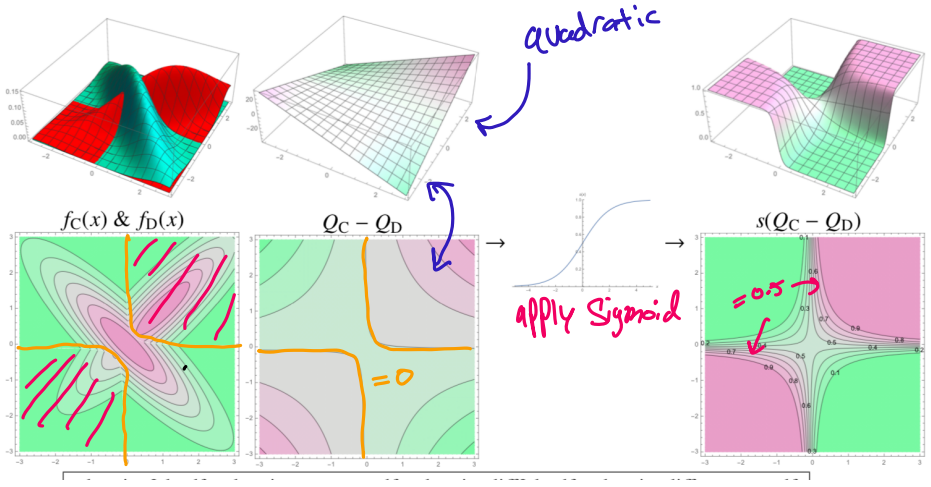
↑ not necessarily isotropic

for 2-classes: Decision fn $Q_C(x) - Q_D(x)$ is quadratic, might be indefinite

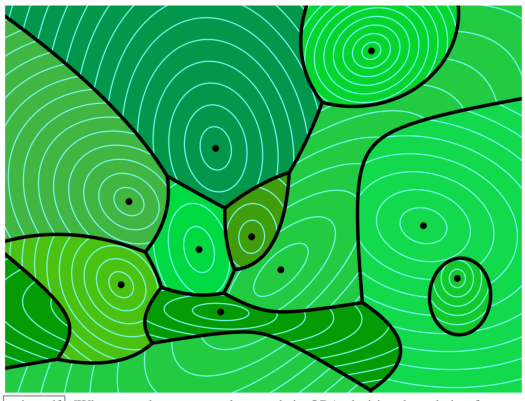
$\Rightarrow$ Bayes decision boundary is quadric

- Posterior is $P(Y=c|X=x) = s(Q_C(x) - Q_D(x))$, $s(\cdot)$ is logistic fn

Solution to quadratic in an arbitrary space



Anisotropic V. diagram



aniso.pdf [When you have many classes, their QDA decision boundaries form an anisotropic Voronoi diagram. Interestingly, a cell of this diagram might not be connected.]

w/ LDA, quadratic terms cancel out

LDA: one $\hat{\Sigma}$ for all classes

$$Q_C(x) - Q_D(x) = (\mu_C - \mu_D)^T \Sigma^{-1} x - \frac{\mu_C^T \Sigma^{-1} \mu_C - \mu_D^T \Sigma^{-1} \mu_D}{2} + \ln \pi_C - \ln \pi_D$$

$w^T x$ $\qquad$ α

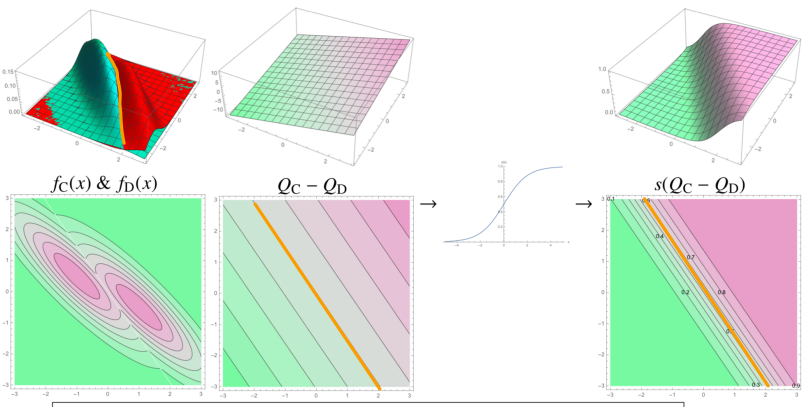
Choose class that maximizes linear discriminant fn:

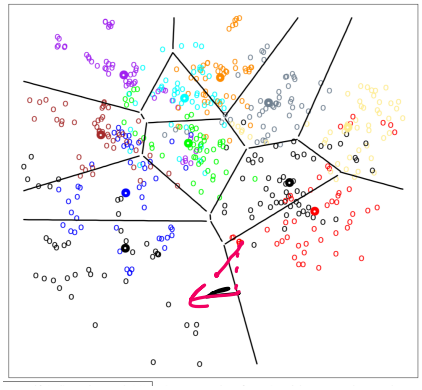
$$\mu_C^T \Sigma^{-1} \vec{x} \qquad -\frac{1}{2} \mu_C^T \Sigma^{-1} \mu_C + \ln \pi_C$$

expensive, once for each class

Compute only 1 discriminant fn, use on all classes

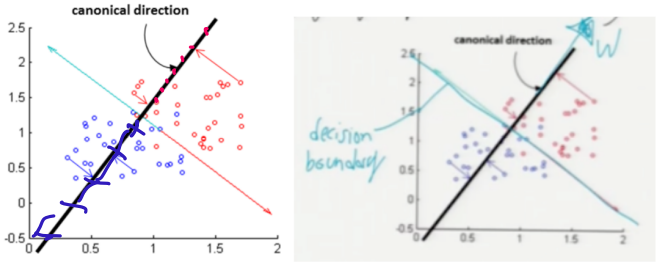
2-classes: - Decision boundary is $w^T x + \alpha = 0$
- Posterior is $P(Y=c \mid X=x) = s(w^T x + \alpha)$





Notes: LDA interpreted as projecting points onto normal vector, cut projection in half → decision boundary

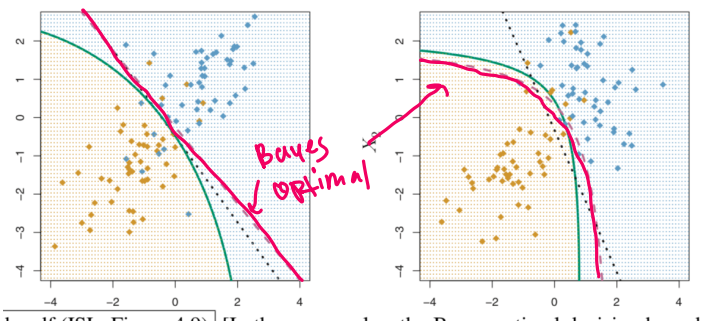
– LDA often interpreted as projecting points onto normal w; cutting the line in half.



project.png

For 2 classes, LDA learns d+1 params ($\vec{w}$ & λ)

- QDA has $\frac{d(d+3)}{2}+1$ parameters (cross terms)
- QDA more likely to overfit to data
- LDA underfits



high

- w/ added features, LDA can be nonlinear

QDA can give boundaries not quadratic

- We don't get true Bayes classifier unless data is truly Gaussian
  - Since we estimate Σ, μ from finite data
  - most datasets aren't perfectly Gaussian
- changing priors or loss = adding constants to discriminant fns
  ↳ change isotopes
- Posterior prob gives decision boundary for 10% prob, 50%, 90%.
  ↳ can show choosing isovalue = prob. p same as choosing asymmetrical loss p for false positive, 1-p for false neg. OR choosing prior to be 1-p; $\pi_0 = p$

Some terms over last few weeks:

$X \in \mathbb{R}^{n \times d}$ is a design matrix of data points, each col. is feature

Centering X: Subtract μ from each row of X $\rightarrow$ μ of each row becomes 0

Let R be uniform. Sample Covariance Matrix: $Var(R) = \frac{1}{n} \dot{X}^T \dot{X}$

Class specific cov. matrix: $\hat{\Sigma}_c = \frac{1}{n_c} X_c^T X_c$

↑ only points of design matrix w/ C

Decorrelating X: eigendecompose Σ, rotate X until it's axis-aligned

$Z = \ddot{X} V$, $Var(R) = V \Lambda V^T$

$\rightarrow Var(Z) = \Lambda$, Z has diag. covariance