Gaussian Discriminant Analysis

Fundamental assumption: each class comes from a normal/gaussian distribution

$X \sim N(\mu, \sigma^2)$: $f(\vec{x}) = \frac{1}{(\sqrt{2\pi}\,\sigma)^d} e^{\frac{-1}{2\sigma^2} \cdot \|\vec{x} - \vec{\mu}\|^2}$ (multivariate form, $\underbrace{\text{variance is same in every direction}}_{\text{isotropic}}$)

- For each class C, suppose we estimate μ_c, σ_c^2, prior $\pi_c = P(Y=c)$.
- then, given x, find Bayes decision rule $r^*(x)$ that predicts class C that maximizes posterior probability which is the same as maximizing $f(X=x \mid Y=c)\,\pi_c \cdot L(z, y)$

add this if not 0-1 loss func.

Want to get rid of exponential:

observe that $\ln(w)$ is monotonically increasing $\forall w > 0$

$\implies$ same as maximizing:

$$Q_c(x) = \ln\left((\sqrt{2\pi})^d \, f_c(x)\, \pi_c\right) = \underbrace{\frac{-\|x - \mu_c\|^2}{2\sigma_c^2}}_{\text{quadratic in } x} - d \ln \sigma_c + \ln \pi_c$$

"Q" = quadratic

normal PDF, estimates $f(X=x \mid Y=c)$

only need to calculate constants once

if 5 classes, compute all 5 $Q_{c_i}(x)$, pick biggest probability (most likely class)

Quadratic Discriminant Analysis (QDA):

Suppose only classes C & D, then

quadratic decision fn.

$$r^*(x) = \begin{cases} C & \text{if } \boxed{Q_C(x) > Q_D(x)} \\ D & \text{o/w} \end{cases}$$

Same answer as computing posterior probabilities

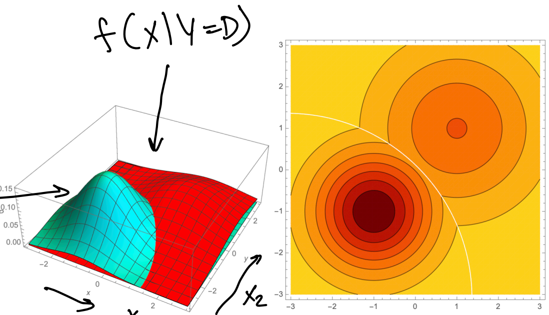
Decision fn is quadratic in X

→ Bayes decision boundary $= \{ x : Q_C(x) > Q_D(x) \}$

- In 1D, B.d.b may have 1 or 2 points
- In d-D, B.d.b is a quadric

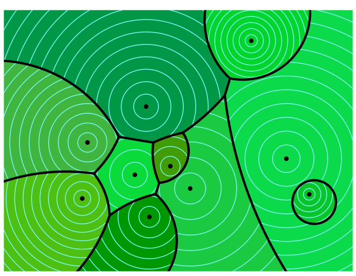
$f(x \mid Y=D)$

$f(x \mid Y=c)$ Gaussian for class C



x_1 x_2

multiple classes, each black dot is a mean of a gaussian distribution



What if we want posterior probabilities?

To get posterior probabilities in 2-class case,

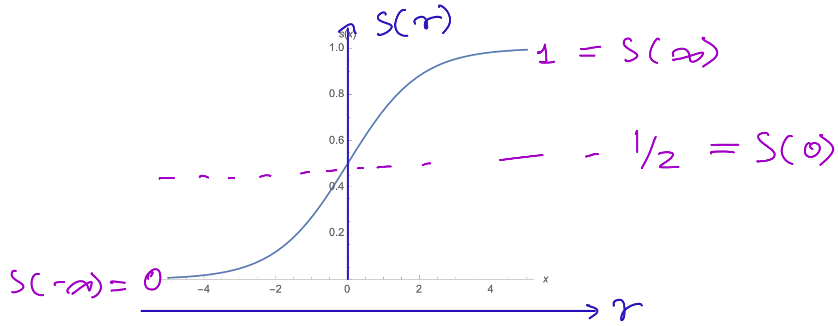
- apply Bayes thm to find $P(Y=c \mid X) = \dfrac{f(X \mid Y=c)\,\pi_c}{f(X \mid Y=c)\,\pi_c + f(X \mid Y=D)\,\pi_D}$

need to relate $f(X \mid Y=c)\,\pi_c$ back to $Q_c(X)$:

- recall $e^{Q_c(x)} = (\sqrt{2\pi})^d f_c(x)\,\pi_d$

$$\Longrightarrow P(Y=c \mid X) = \frac{e^{Q_c(x)}}{e^{Q_c(x)} + e^{Q_d(x)}} = \frac{1}{1+e^{Q_D(x)-Q_c(x)}}$$

SIGMOID fn: $= s(Q_c(x) - Q_D(x))$, where $s(\gamma) = \dfrac{1}{1+e^{-\gamma}}$



Linear Discriminant Analysis (LDA):

- every class has same variance/covariance ← key assumption $\sigma_c = \sigma_D$
  - ↳ always linear decision boundaries
    - ↳ less likely to overfit

only dependence on x, linear term

$$\longrightarrow Q_c(X) - Q_D(X) = \underbrace{\frac{(\mu_c - \mu_D)\cdot X}{\sigma^2}}_{W \cdot X} \underbrace{- \frac{\|\mu_c\|^2 - \|\mu_D\|^2}{2\sigma^2} + \ln \pi_c + \ln \pi_D}_{\alpha}$$

If multiple classes, compute LD fn for each class, then choose C that maximizes those LD fns.

LD fn: $\dfrac{\mu_c \cdot x}{\sigma^2} - \dfrac{\|\mu_c\|^2}{2\sigma^2} + \ln \pi_c$

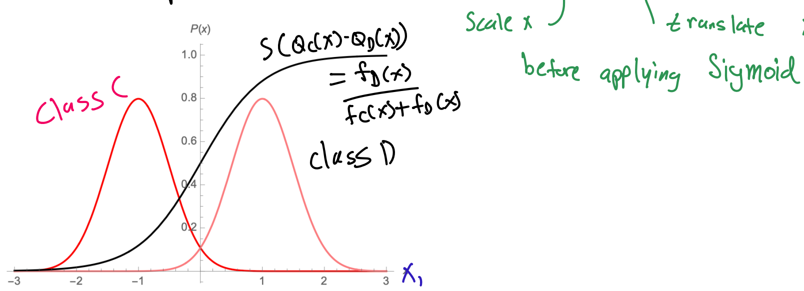
Same as doing $Q_C(x) - Q_D(x)$ for every possible pair

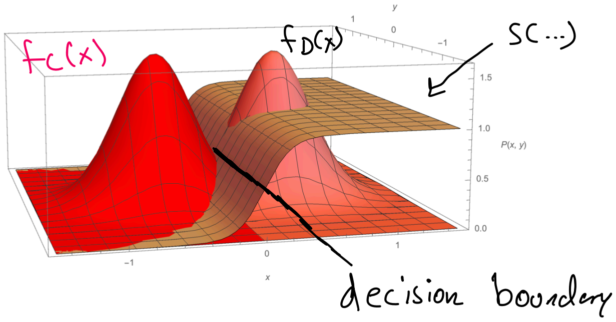
e.g. MNIST, calculate $Q_0(x) \ldots Q_9(x)$ and pick biggest

In 2-class case: - decision boundary $= \underbrace{W \cdot X + \alpha}_{} = 0$ → $W = \dfrac{\mu_c - \mu_D}{\sigma^2}$, α above

- posterior prob $P(Y=c \mid X=x) = s(W \cdot x + \alpha)$

scale x ↑ translate x before applying sigmoid





If $\pi_C = \pi_D = \frac{1}{2}$, then $\underbrace{(\mu_C - \mu_D)\cdot x - (\mu_C - \mu_D)\cdot \frac{(\mu_C + \mu_D)}{2} = 0}_{\text{decision boundary}}$

This is the centroid method

MAXIMUM LIKELIHOOD estimation of parameters

(Ronald Fisher, 1912)

Let's flip a coin ~ Bernoulli(P)

10 flips, 8 heads, 2 tails

What value of p gives this outcome?

- let X = # heads, $X \sim \text{Binom}(n, p)$

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

find p that maximizes this

$$\Rightarrow P(X=8) = \binom{10}{8} p^8 (1-p)^2 = 45 p^8 (1-p)^2 = \mathcal{L}(p)$$ (likelihood fn)

Probability of 8 heads in 10 flips:
written as $\mathcal{L}(p)$ of distr. params, this is the likelihood func.

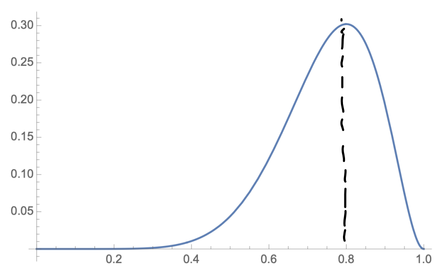
Maximum likelihood estimation (MLE): method of estimating params of a statistical distribution by picking params that maximize $\mathcal{L}$
— one method of density estimation: estimating PDF from data

→ Find p that maximizes $\mathcal{L}(p)$:

Solve by setting $\frac{d}{dp}$ to 0:

$$\frac{d}{dp}\mathcal{L}(p) = 360p^7(1-p)^2 - 90p^8(1-p) = 0$$

$\rightarrow 4(1-p) - p = 0 \Rightarrow p = 0.8$



$\hat{\pi}_C = 0.8$ = chance a newly sampled point is in class C

MLE used on a Gaussian:

given $X_1 \dots X_n$, find best-fit Gaussian

Since this is continuous, the likelihood is no longer a probability

regardless, likelihood of generating these points is:

$$\mathcal{L}(\mu, \sigma; x_1 \dots x_n) = f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n)$$ (intersection)

Choose μ & σ that maximize $\mathcal{L}$

take log to get a sum:

log likelihood: $\ell(\mu, \sigma; x_1 \dots x_n)$ is natural log of $\mathcal{L}(\cdot)$

maximizing likelihood $\Longleftrightarrow$ maximizing log likelihood

$$\longrightarrow \ell(\cdot) = \ln f(x_1) + \dots + \ln f(x_n)$$

$$= \sum_{i=1}^{n} \Big(\underbrace{\frac{-\|x_i - \mu\|^2}{2\sigma^2} - d \ln \sqrt{2\pi} - d \ln \sigma}_{\ln \text{ of normal PDF}} \Big)$$

Want to set $\nabla_\mu \ell = 0$, $\frac{\partial L}{\partial \sigma} = 0$ to find critical point

$$\nabla_\mu \mathcal{L} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(average of samples is good est. for μ of distribution)

mean estimate of true μ

$$\frac{\partial \ell}{\partial \sigma} = \sum_{i=1}^{n} \frac{\|x_i - \mu\|^2 - d\sigma^2}{\sigma^3} \implies \underbrace{\hat{\sigma}^2 = \frac{1}{dn} \sum_{i=1}^{n} \|x_i - \hat{\mu}\|^2}_{\text{Sample Variance}}$$

don't know exact μ