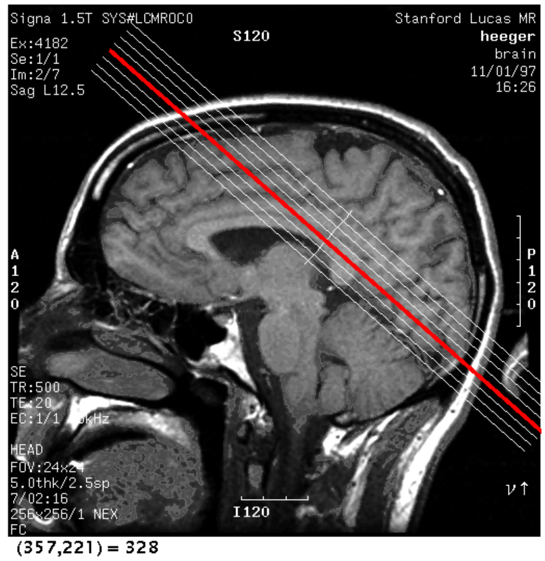
Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2023

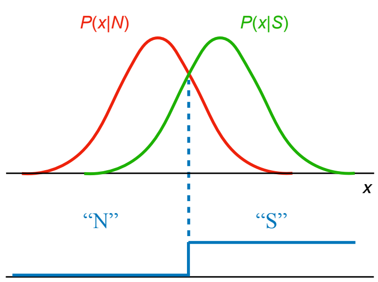
Section 5a:

Statistical Decision Theory
+
Signal Detection Theory



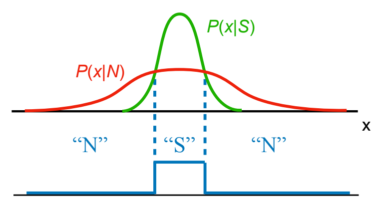
Tumor, or not?

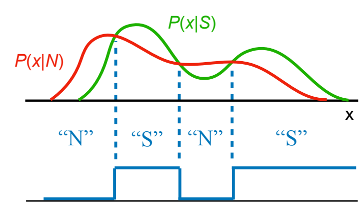
Signal Detection Theory (binary estimation)



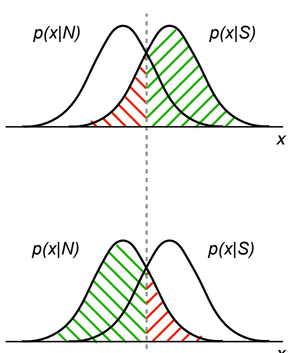
For equal-shape, unimodal, symmetric distributions, the ML decision rule is a *threshold* function.

More generally, decision rule can have multiple thresholds…

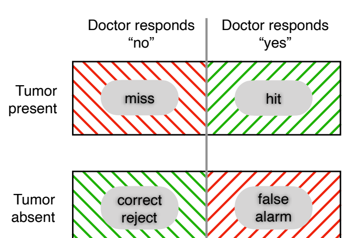




Signal Detection Theory: Potential outcomes



decision
threshold or
criterion *c*

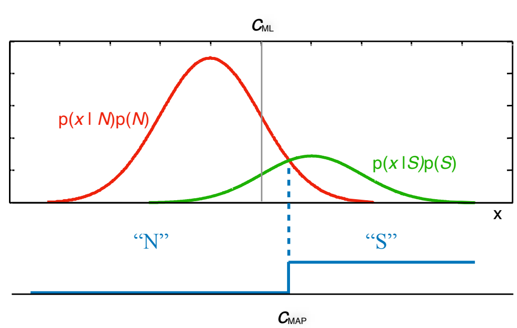


For criterion *c*, and cumulative distribution functions *P*, e.g., $P(c \mid S) = p(x \leq c \mid S)$

$P(c \mid S)$ = p(miss)
$1 - P(c \mid S)$ = p(hit)
$P(c \mid N)$ = p(correct reject)
$1 - P(c \mid N)$ = p(false alarm)

MAP decision rule?

MAP solution maximizes proportion of correct answers, *taking prior probability into account*.



Compared to ML threshold, the MAP criterion moves *away* from higher-probability option.

Bayes decision rule?

Incorporate values for the four possible outcomes:

Payoff Matrix

Stimulus \ Response	Yes	No
S+N	V_{S+N}^{Yes}	V_{S+N}^{No}
N	V_{N}^{Yes}	V_{N}^{No}

Bayes Optimal Criterion

Stimulus \ Response	Yes	No
S+N	V_{S+N}^{Yes}	V_{S+N}^{No}
N	V_{N}^{Yes}	V_{N}^{No}

$$\mathbb{E}(Yes \mid x) = V_{S+N}^{Yes} p(S+N \mid x) + V_{N}^{Yes} p(N \mid x)$$
$$\mathbb{E}(No \mid x) = V_{S+N}^{No} p(S+N \mid x) + V_{N}^{No} p(N \mid x)$$

Say yes if $\mathbb{E}(Yes \mid x) \geq \mathbb{E}(No \mid x)$

Optimal Criterion

$$\mathbb{E}(Yes \mid x) = V_{S+N}^{Yes} \mathrm{p}(S+N \mid x) + V_{N}^{Yes} \mathrm{p}(N \mid x)$$
$$\mathbb{E}(No \mid x) = V_{S+N}^{No} \mathrm{p}(S+N \mid x) + V_{N}^{No} \mathrm{p}(N \mid x)$$

Say yes if $\mathbb{E}(Yes \mid x) \geq \mathbb{E}(No \mid x)$

Say yes if $\dfrac{\mathrm{p}(S+N \mid x)}{\mathrm{p}(N \mid x)} \geq \dfrac{V_N^{No} - V_N^{Yes}}{V_{S+N}^{Yes} - V_{S+N}^{No}} = \dfrac{V(\text{Correct} \mid N)}{V(\text{Correct} \mid S+N)}$

Posterior odds

Apply Bayes' Rule

Posterior — Likelihood — Prior

$$p(S+N \mid x) = \frac{p(x \mid S+N)\,p(S+N)}{p(x)}$$

Nuisance normalizing term

$p(N \mid x) = \dfrac{p(x \mid N)\,p(N)}{p(x)}$, hence

$$\frac{p(S+N \mid x)}{p(N \mid x)} = \left(\frac{p(x \mid S+N)}{p(x \mid N)}\right)\left(\frac{p(S+N)}{p(N)}\right)$$

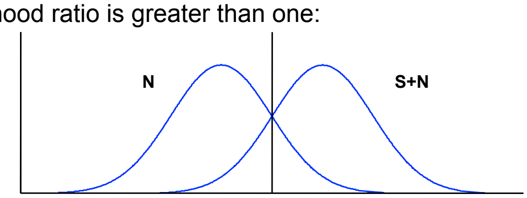
Posterior odds — Likelihood ratio — Prior odds

Optimal Criterion

Say yes if $\dfrac{p(S+N \mid x)}{p(N \mid x)} \geq \dfrac{V(\text{Correct} \mid N)}{V(\text{Correct} \mid S+N)}$

i.e., if $\dfrac{p(x \mid S+N)}{p(x \mid N)} \geq \dfrac{p(N)}{p(S+N)} \dfrac{V(\text{Correct} \mid N)}{V(\text{Correct} \mid S+N)} = \beta$

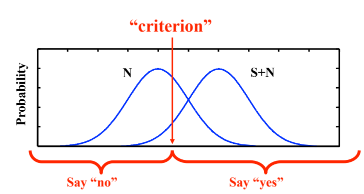
Example, if equal priors and equal payoffs, say yes if the likelihood ratio is greater than one:



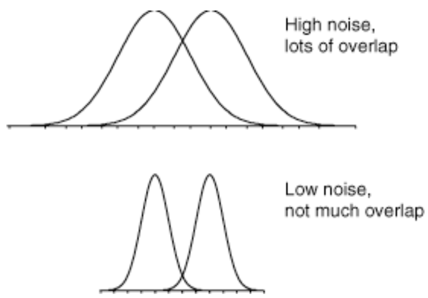
Example applications of SDT

- Vision
  - Detection (something vs. nothing)
  - Discrimination (lower vs greater level of: intensity, contrast, depth, slant, size, frequency, loudness, ...
- Memory (internal response = trace strength = familiarity)
- Neurometric function/discrimination by neurons (internal response = spike count)

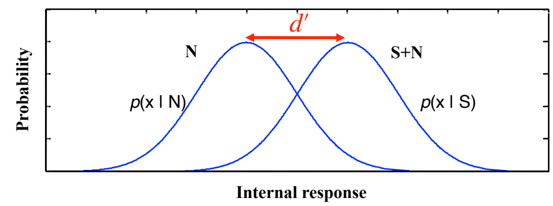
From experimental measurements, assuming human is optimal, can we determine the underlying distributions and criterion?



Signal Detection Theory: discriminability (d')



Internal response: probability of occurrence curves

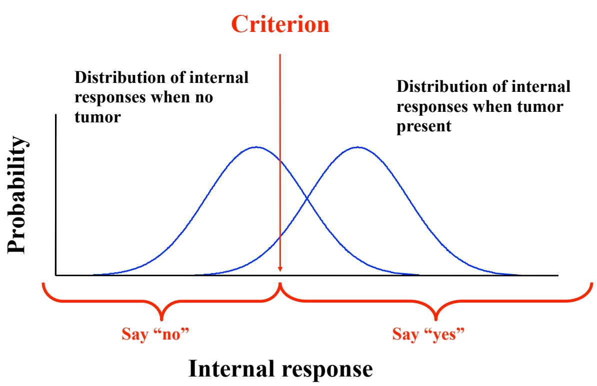


$$d' = \frac{\text{“separation”}}{\text{“width”}}$$

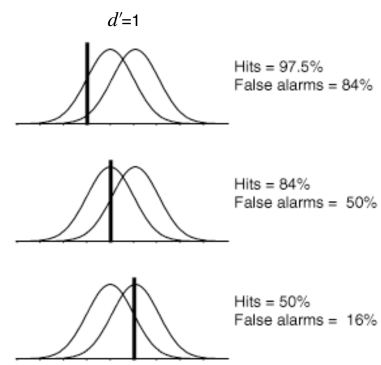
Discriminability (“d-prime”) is the normalized separation between the two distributions

Error rate is a function of d'

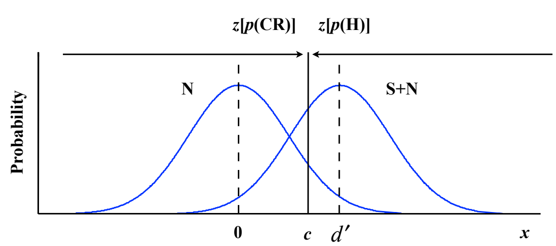
Criterion



Signal Detection Theory: Criterion



SDT: Gaussian case

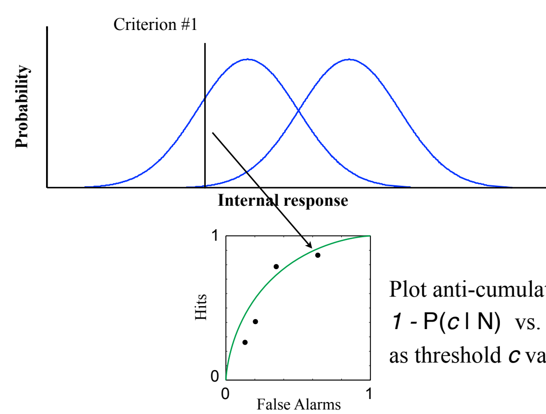


$$d' = z[\mathrm{p}(H)] + z[\mathrm{p}(CR)] = z[\mathrm{p}(H)] - z[\mathrm{p}(FA)]$$

$$\mathrm{c} = z[p(CR)] \qquad G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

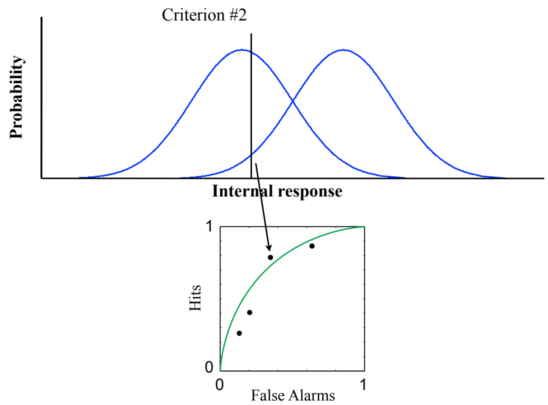
$$\beta = \frac{p(x=c\,|\,S+N)}{p(x=c\,|\,N)} = \frac{e^{-(c-d')^2/2}}{e^{-c^2/2}} \qquad (\text{Fix } \sigma = 1)$$

ROC (Receiver Operating Characteristic)

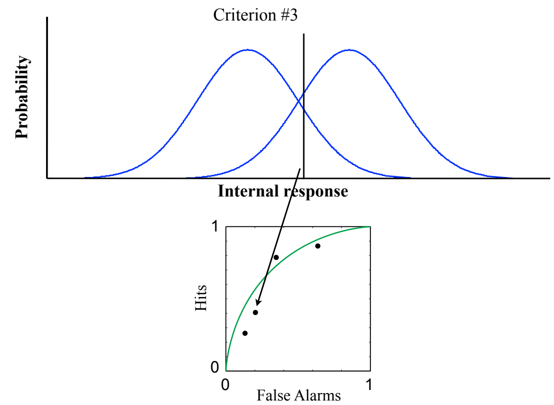


Plot anti-cumulatives:
1 - P(c | N) vs. 1 - P(c | S)
as threshold c varies

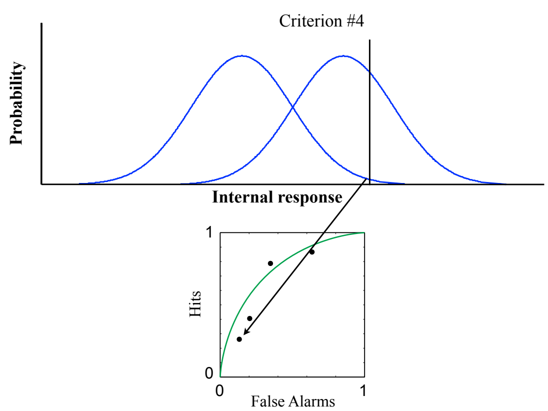
ROC (Receiver Operating Characteristic)



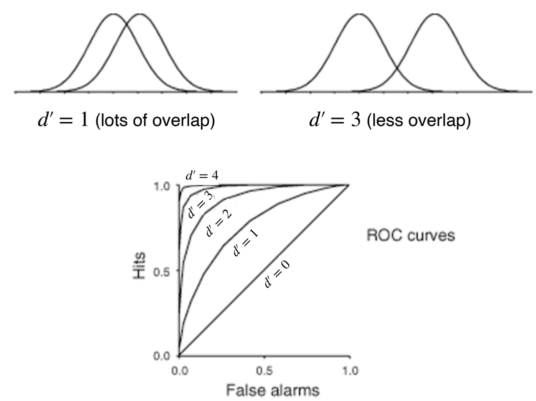
ROC (Receiver Operating Characteristic)



ROC (Receiver Operating Characteristic)



ROC (Receiver Operating Characteristic)



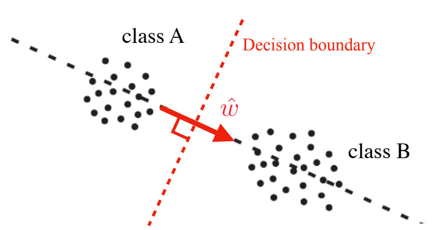
[on board: Area under curve = %correct in a 2AFC task]

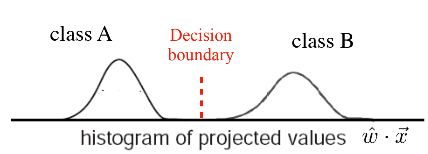
Decision/classification in multiple dimensions

- Data-driven linear classifiers:
  - Prototype Classifier - minimize distance to class mean
  - Fisher Linear Discriminant (FLD) - maximize d'
  - Support Vector Machine (SVM) - maximize margin
- Statistical:
  - ML/MAP/Bayes under a probabilistic model
  - e.g.: Gaussian, identity covariance (same as Prototype)
  - e.g.: Gaussian, equal covariance (same as FLD)
  - e.g.: Gaussian, general case (Quadratic Discriminator)
- Some Examples:
  - Visual gender classification
  - Neural population decoding

Linear Classifier

Find unit vector $\hat{w}$ ("discriminant") that best separates the distributions

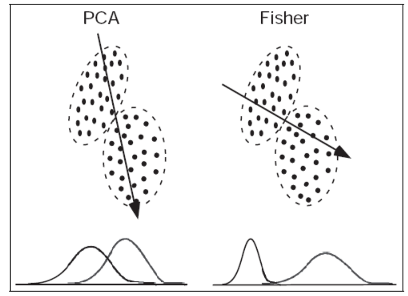




Simplest linear discriminant: the Prototype Classifier

$$\hat{w} = \frac{\vec{\mu}_A - \vec{\mu}_B}{\|\vec{\mu}_A - \vec{\mu}_B\|}$$

Fisher Linear Discriminant



$$\max_{\hat{w}} \frac{\left[\hat{w}^T(\vec{u}_A - \vec{u}_B)\right]^2}{\left[\hat{w}^T C_A \hat{w} + \hat{w}^T C_B \hat{w}\right]}$$ (note: this is d' squared!)

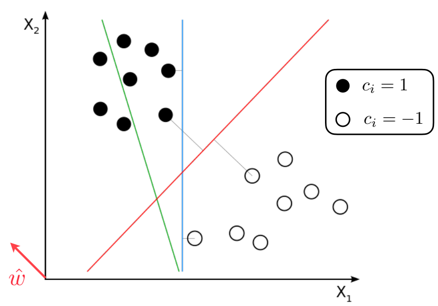
optimum: $\hat{w} = C^{-1}(\vec{u}_A - \vec{u}_B)$, where $C = \frac{1}{2}(C_A + C_B)$

Support Vector Machine (SVM)

(widely used in machine learning, but no closed form solution)

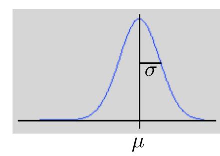
Maximize the "margin" (gap between data sets):

find largest m, and $\{\hat{w}, b\}$ s.t. $c_i(\hat{w}^T \vec{x}_i - b) \geq m, \quad \forall\, i$

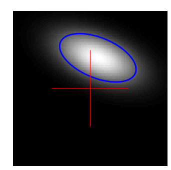


Reminder: Multi-D Gaussian densities

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



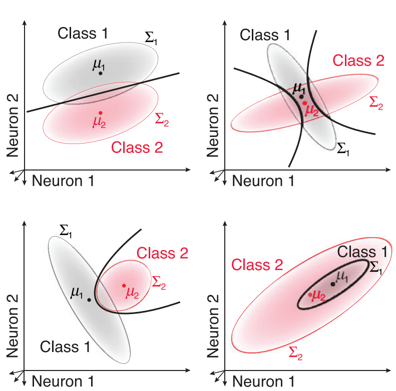
$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$



mean: [0.2, 0.8]
cov: [1.0 -0.3;
-0.3 0.4]

ML (or MAP) classifier for two Gaussians

Decision boundary is *quadratic*, with four possible geometries:



[on board]

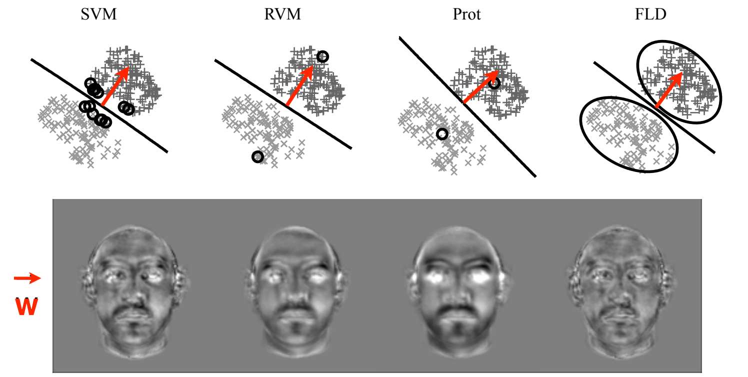
[figure: Pagan et al. 2016]

A perceptual example: Gender identification

- 200 face images (100 male, 100 female)
- Adjusted for position, size, intensity/contrast
- Labeled by 27 human subjects

[Graf & Wichmann, NIPS*03]

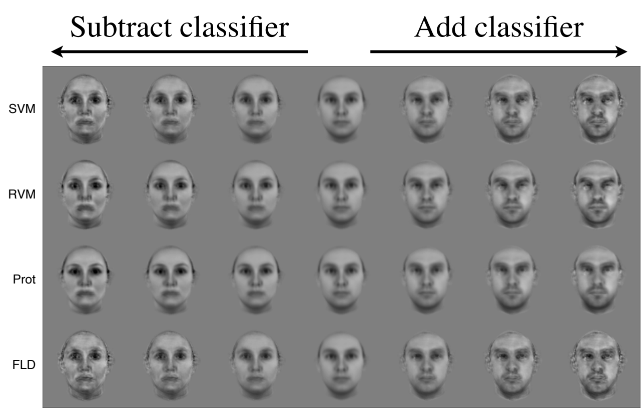
Linear classifiers



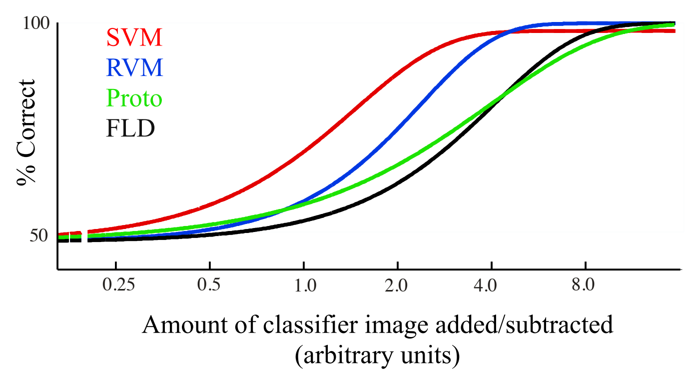
Four linear classifiers trained on subject data

Model validation/testing

- Cross-validation: Subject responses [% correct, reaction time, confidence] are explained
  - very well by SVM
  - moderately well by RVM / FLD
  - not so well by Prot
- Curse of dimensionality strongly limits this result. A more direct test: Synthesize optimally discriminable faces...



[Wichmann, et. al; NIPS*04]



[Wichmann, et. al; NIPS*04]

Fisher Information

- Second-order expansion of the (expected) negative log likelihood:

$$I(s) = -\mathbb{E}\left[\frac{\partial^2 \log p(r|s)}{\partial s^2}\right]$$

- Provides a bound on "precision" of unbiased estimators: (the "Cramer-Rao" bound) $\sigma^2(s) \geq \dfrac{1}{I(s)}$

- Perceptually, provides a bound on discriminability: (Series et. al. 2009) $D(s) \leq \sqrt{I(s)}$

- Examples: with mean stimulus response *f(s)*

Gaussian case: $p(r|s) \sim \mathcal{N}(f(s), \sigma^2) \qquad I(s) = [f'(s)]^2/\sigma^2$

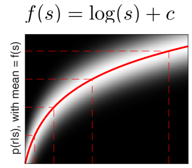
Poisson case: $p(r|s) \sim \text{Poiss}(f(s)) \qquad I(s) = [f'(s)]^2/f(s)$

Example: Weber's law

$$D(s) \propto \frac{1}{s} \leq \sqrt{I(s)}$$

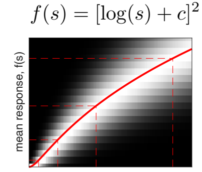
Assuming $I(s) \propto \dfrac{1}{s^2}$ what internal representation explains this? Many!

additive Gaussian noise, with mean $f(s) = \log(s) + c$



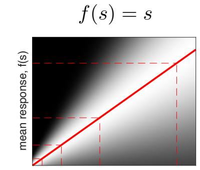
entirely due to response mean (Fechner, 1860)

Poisson noise, with mean $f(s) = [\log(s) + c]^2$



discrete representation, depends on both mean and variance

multiplicative Gaussian noise, with mean $f(s) = s$



entirely due to response variance