Lecture 9: Linear Classification

Optimal f^* is

(1) regression: $L(y, f(x)) = (y - f(x))^2$

$$f^*(\underset{\sim}{x}) = \mathbb{E}[Y \mid \underset{\sim}{X} = \underset{\sim}{x}]$$

(2) classification: $L(y, f(\underset{\sim}{x})) = \mathbb{1}(y \neq f(\underset{\sim}{x}))$

$$f^*(\underset{\sim}{x}) = \arg\max_c \underline{\mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})}$$

Calculate / approx. $\mathbb{P}(Y = c_1 \mid \underset{\sim}{X} = \underset{\sim}{x}), \mathbb{P}(Y = c_2 \mid \underset{\sim}{X} = \underset{\sim}{x}), \ldots$

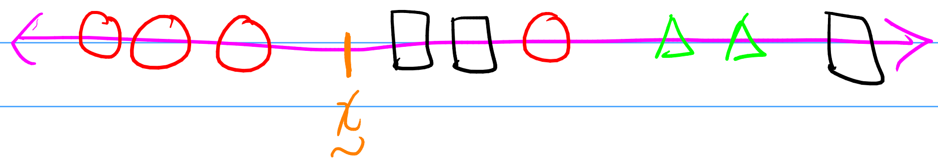
↑ pick class w/ largest prob.

Simple example: KNN classification

$$\mathbb{P}(Y = c_i \mid \underset{\sim}{X} = \underset{\sim}{x}) \approx$$ of points near $\underset{\sim}{x}$ what pct. are in class c_i $\;\; = \hat{P}_i$

$$\hat{f}(\underset{\sim}{x}) = \arg\max_i \hat{P}_i$$

= predict as class w/ most of nearby points



$k=3$ classifier then predict $\square$

$$\hat{p}_i = \text{pct of neighbors in class } c_i = \frac{1}{k} \underbrace{\sum_{x_n \in N_k(\underset{\sim}{x})} \mathbb{1}(y_n = c_i)}_{\text{\# of neighbors in class } c_i}$$

Bayes' Classifier says to look at $P(Y=c \mid \underset{\sim}{X} = \underset{\sim}{x})$

1. discriminative models

   $\rightarrow$ estimate/model $\mathbb{P}(Y=c \mid \underset{\sim}{X} = \underset{\sim}{x})$ directly

   i.e. model $Y \mid \underset{\sim}{X} = \underset{\sim}{x}$

   $\hookrightarrow$ kNN, logistic regression, classification trees

2. generative models

   Bayes' rule: $$\mathbb{P}(Y=c \mid \underset{\sim}{X} = \underset{\sim}{x}) = \frac{\mathbb{P}(\underset{\sim}{X} = \underset{\sim}{x} \mid Y=c)\,\mathbb{P}(Y=c)}{\mathbb{P}(\underset{\sim}{X} = \underset{\sim}{x})}$$

   $$\propto \mathbb{P}(\underset{\sim}{X} = \underset{\sim}{x} \mid Y=c)\,\mathbb{P}(Y=c)$$

   model $\underset{\sim}{X} \mid Y$ and Y

   $\hookrightarrow$ LDA/QDA, naive bayes

Defn: Linear Classifier

Bayes': $\hat{f}(\underset{\sim}{x}) = \underset{c}{\operatorname{argmax}}\ \mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})$

more generally

$$\hat{f}(\underset{\sim}{x}) = \underset{c}{\operatorname{argmax}}\ \delta_c(\underset{\sim}{x})$$

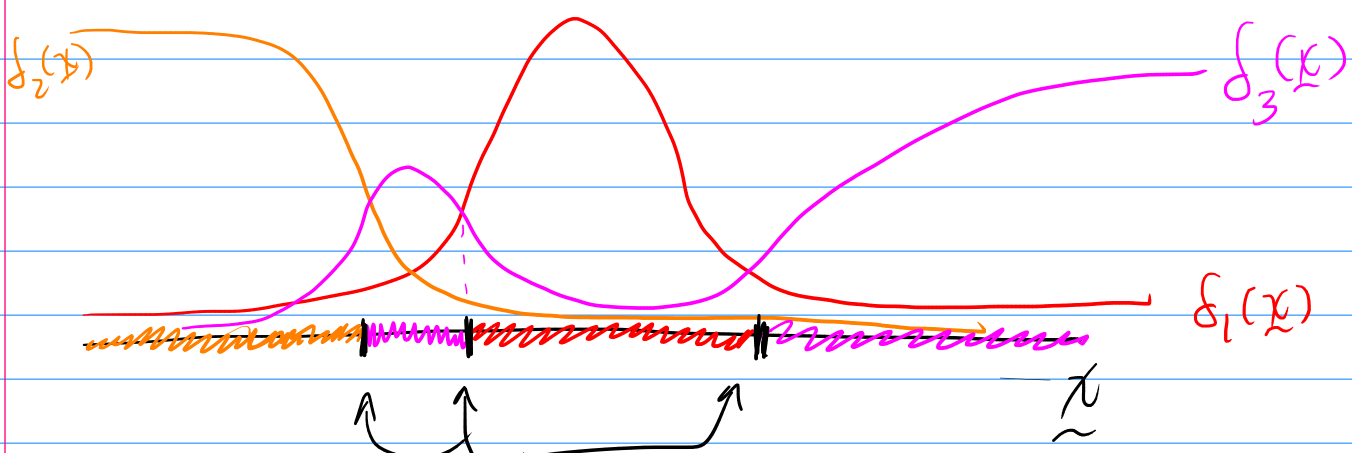
discriminant functions

Large if likely class c
small otherwise

Ex, $\delta_c(\underset{\sim}{x}) = \mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})$

Ex, $\delta_c(\underset{\sim}{x}) = \log \mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})$

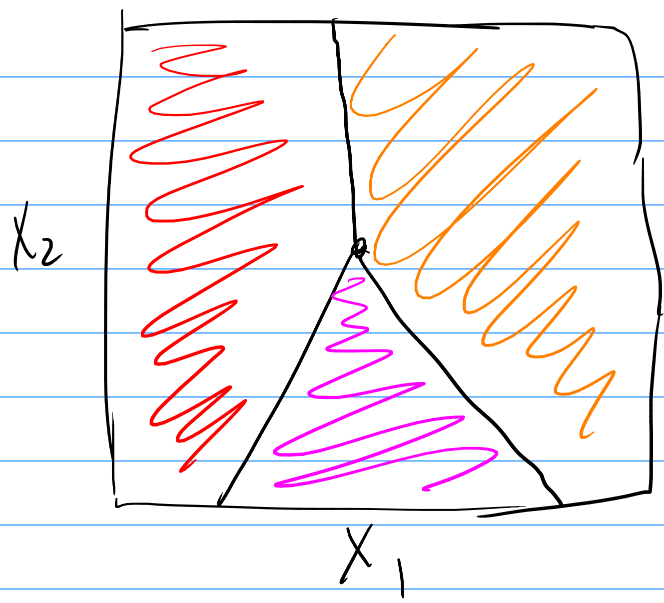
Ex, $\delta_c(\underset{\sim}{x}) = \alpha_c + \beta_c^T \underset{\sim}{x}$



decision boundaries
class boundaries

non-linear boundaries

linear boundaries



Defn: Linear classifier has linear decision boundaries

Defn: Linear Classifier a classifier whose discriminant functions can be transformed to linear functions of $\underset{\sim}{x}$ by a increasing transformation. $(\delta_c(\underset{\sim}{x}) = \alpha_c + \beta_c^T \underset{\sim}{x})$

Ex, $\delta_c(\underset{\sim}{x}) = \alpha_c + \beta_c^T \underset{\sim}{x}$ then the classifier is linear

Ex. increasing function T so that

$$T(\delta_c(\underset{\sim}{x})) = \alpha_c + \beta_c^T \underset{\sim}{x}$$

is a linear classifier

Two class problem classes 1 and 2

decision boundary:

$$\{\underset{\sim}{x} : \delta_1(\underset{\sim}{x}) = \delta_2(\underset{\sim}{x})\}$$

So if δ_c are linear then

$$\delta_1(\underset{\sim}{x}) = \boxed{\alpha_1 + \beta_1^T \underset{\sim}{x} = \alpha_2 + \beta_2^T \underset{\sim}{x}} = \delta_2(\underset{\sim}{x})$$

$$\Rightarrow (\beta_1 - \beta_2)^T \underset{\sim}{x} = \alpha_2 - \alpha_1 \qquad [Ax = b]$$

a linear system

Solutions are same (linear) subspace

So the decision boundary is a linear subspace.

Generally, if $\exists$ T to "linearize" δ_c then

decision boundary: maybe not linear but if I apply T

$$\delta_1(\underset{\sim}{x}) = \delta_2(\underset{\sim}{x})$$

$$\Rightarrow T(\delta_1(\underset{\sim}{x})) = T(\delta_2(\underset{\sim}{x}))$$

now both sides are linear

So my boundaries are a subspace

Applying T gives the <u>same classifier</u>

$$\hat{f}(\underset{\sim}{x}) = \arg\max_{c} \delta_c(\underset{\sim}{x})$$

$$= \arg\max_{c} T(\delta_c(\underset{\sim}{x}))$$

inc. fn.