STAT 535

10/6/22

ZOOM

Lecture 3

- Next Lecture : in person
- HW1 – t.b. posted TODAY
- 12-oct 4 } t.b. posted –"–
  13-oct 6
  ↳ on web : handouts
- Q1 → Tue oct 18
- Steve WR : Tutorial Refresher

- Classification specific concepts
- Kinds of predictors
  Nearest Neighbors

Lecture Notes I – Examples of Predictors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

September 29, 2022

Prediction problems by the type of output ✓

The "learning" paradigm and vocabulary ✓ ⊃ Classification concepts

The Nearest-Neigbor and kernel predictors ←
?

Linear predictors

TUTORIAL

- Least squares regression
- Linear Discriminant Analysis (LDA)
- QDA (Quadratic Discriminant Analysis) Not Linear
- Logistic Regression
- The PERCEPTRON algorithm

Classification and regression tree(s) (CART)

The Naive Bayes classifier

Hastie & al

Reading HTF Ch.: 2.3.1 Linear regression, 2.3.2 Nearest neighbor, 4.1–4 Linear classification, 6.1–3. Kernel regression, 6.6.2 kernel classifiers, 6.6.3 Naive Bayes, 9.2 CART, 11.3 Neural networks, Murphy Ch.: 1.4.2 nearest neighbors, 1.4.4 linear regression, 1.4.5 logistic regression, 3.5 and 10.2.1 Naive Bayes,4.2.1–3 linear and quadratic discriminant, 14.7.3– kernel regression, locally weighted regression, 16.2.1–4 CART, (16.5 neural nets), Bach Ch.: N/A

The "sign trick" for transforming a regressor into a classifier

The *sign* function $\mathrm{sgn}(y) = y/|y|$ if $y \neq 0$ and 0 iff $y = 0$ turns a real valued variable Y into a discrete-valued one. This function is used to allow one to construct *real-valued classifiers*. In these classifiers, the model $f(x)$ is a real-valued function, and the prediction $\hat{y}$ is given by $\mathrm{sgn}(f(x))$.

Note that in a vanishingly small fraction of cases, when the value of $f(x)$ is exactly 0, no label will be assigned to the input x.

Classifiers with Continuous output

<u>Binary</u> $y \in \{\pm 1\}$ Classifier $f : \mathcal{X} \to \{\pm 1\}$

space of inputs x

e.g. $\mathbb{R}^d$

But sometimes

Learned

$$f : \mathcal{X} \longrightarrow \mathbb{R}, \quad (f \in \mathcal{F} \text{ model class})$$

$$\operatorname{sgn} z = \begin{cases} +1, & z > 0 \\ -1, & z < 0 \\ 0 & z = 0 \end{cases}$$

$$\hat{y} = \operatorname{sgn} \underbrace{f(x)}_{}$$

classifier output

label assigned by classifier to input x

Data $\mathcal{D} = \{(x^i, y^i),\ i = 1{:}n\}$ $y^i \in \{\pm 1\}$

in our model

Ex: Naive Bayes $\longrightarrow f(x) \in (0,1)$ $f(x) = P_r[Y = 1 \mid X = x] \Rightarrow$

Logistic regression

[Support Vector Machines]

[kernel classifiers] later

[Neural Net]

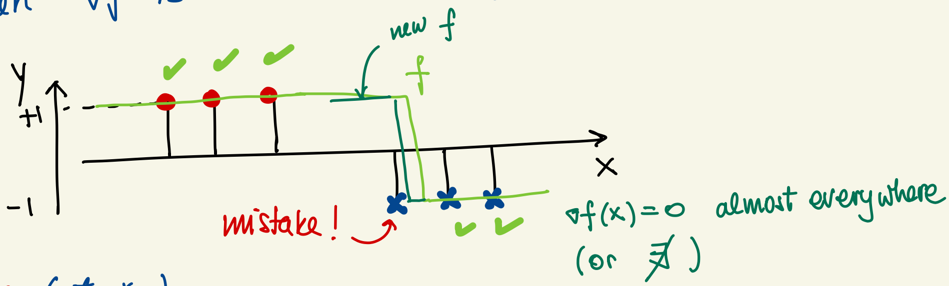
$$\Rightarrow \hat{y}(x) = \operatorname{sgn}\left(f(x) - \frac{1}{2}\right)$$

$$\Leftrightarrow \hat{y}(x) = 1 \text{ iff } P[Y = 1 \mid X = x] > 1/2$$

Why (continuous) $\mathbb{R}$-valued f?

- for statistical models (probabilistic), $f = \Pr[y=1|x]$ (e.g. logistic regression)
- for non-statistical, $|f(x)|$ large ≈ confidence in $\hat{y}$ ← qualitative (e.g. SVM, Neural Net, kernel)
- f differentiable ($\Leftrightarrow \nabla f$ exists) and ∇f not 0 almost everywhere)

then ∇f is used in learning algorithm



Margin of classifier (at x)

- $z = y f(x)$ ⇒ P1 $z > 0$ iff $f(x)$ correct at x

  < 0 iff mistake at x

  (y ↑ true)

P2 $|z| = |f(x)|$ is classifier's confidence

↳ $z > 0$ and large → classifier "knows" it's correct Robustness!!

$z \gtrsim 0$ → Not robust

Continuous valued Multiway $y \in \{1, 2, \dots m\}$

Train $f_{1:m}$ "classifiers"

$$\hat{y}(x) = \underset{c=1:m}{\operatorname{argmax}} f_c(x)$$

Margin $\boxed{z_{(x)} = f_y(x) - \max_{c \neq y} f_c(x)}$

true

correct

$\hat{y} = y \Rightarrow z = f_y - f_{\text{next largest}} > 0$

error

$\hat{y} \neq y \Rightarrow z = f_y - f_{\hat{y}} <$

Decision regions, decision boundary of a classifier

Let $f(x)$ be a classifier (not necessarily binary)

- $f(x)$ $\hat{y}(x)$ takes only a finite set of values
- The **decision region** associated with class y = the region in X space where f takes value y, i.e. $D_y = \{x \in \mathbb{R}^d,\ f(x) = y\} = \boxed{f^{-1}(y)}$.
- The boundaries separating the decision regions are called **decision boundaries**.

some label

any x, y, f

$f : \mathcal{X} \to \mathcal{Y}$

$A \subseteq \mathcal{Y}$

$f^{-1}(A) = \{x \in \mathcal{X} \mid f(x) \in A\}$

Binary

$\hat{y} = +1$

$D_+ = \{x \in \mathbb{R}^d,\ f(x) > 0\}$

$D_- = \{x \in \mathbb{R}^d,\ f(x) < 0\}$

decision boundary $D_0 = \{x \in \mathbb{R}^d,\ f(x) = 0\}$

f real valued

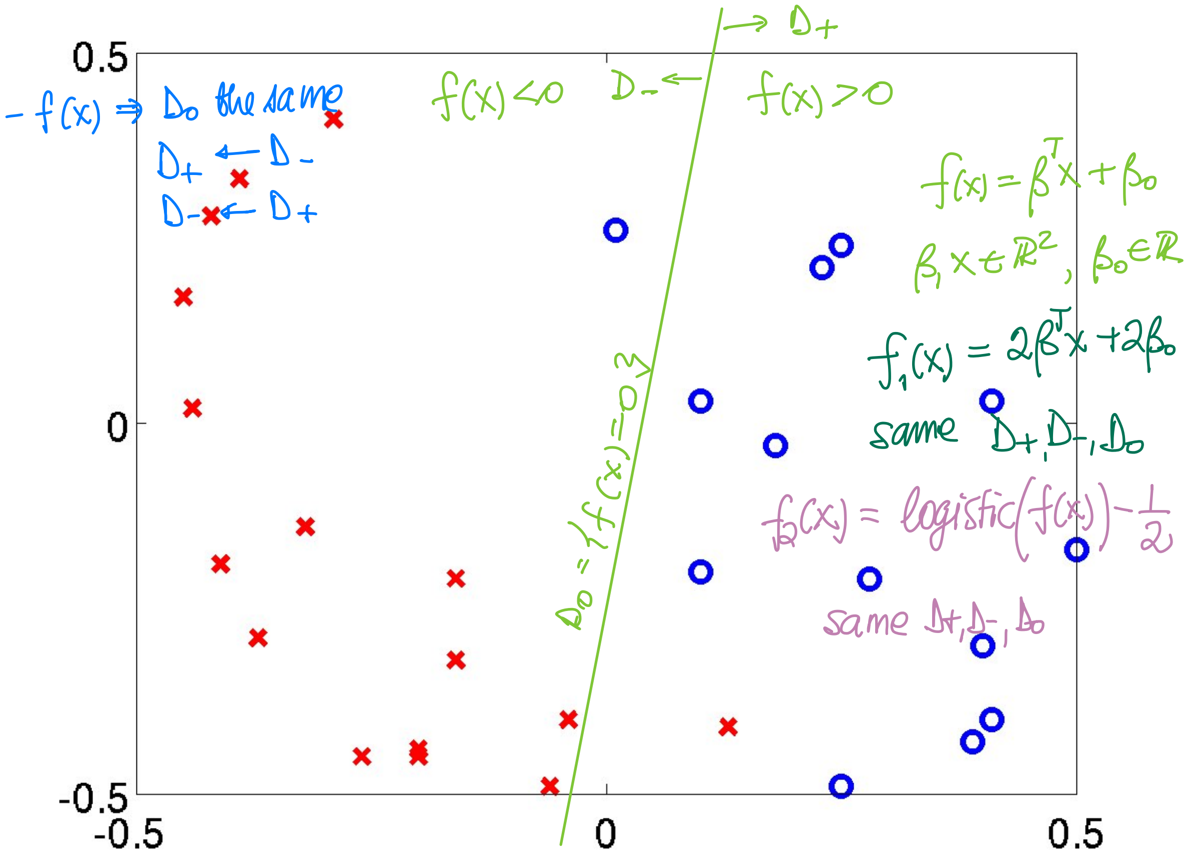
D_0 = boundary between D_+, D_-

Decision regions, decision boundary of a classifier

Let $f(x)$ be a classifier (not necessarily binary)

- $f(x)$ takes only a finite set of values
- The **decision region** associated with class y = the region in X space where f takes value y, i.e. $D_y = \{x \in \mathbb{R}^d, f(x) = y\} = f^{-1}(y)$.
- The boundaries separating the decision regions are called **decision boundaries**.

Multiway D_0 = boundary between any two $D_y, D_{y'}$
$y \neq y'$

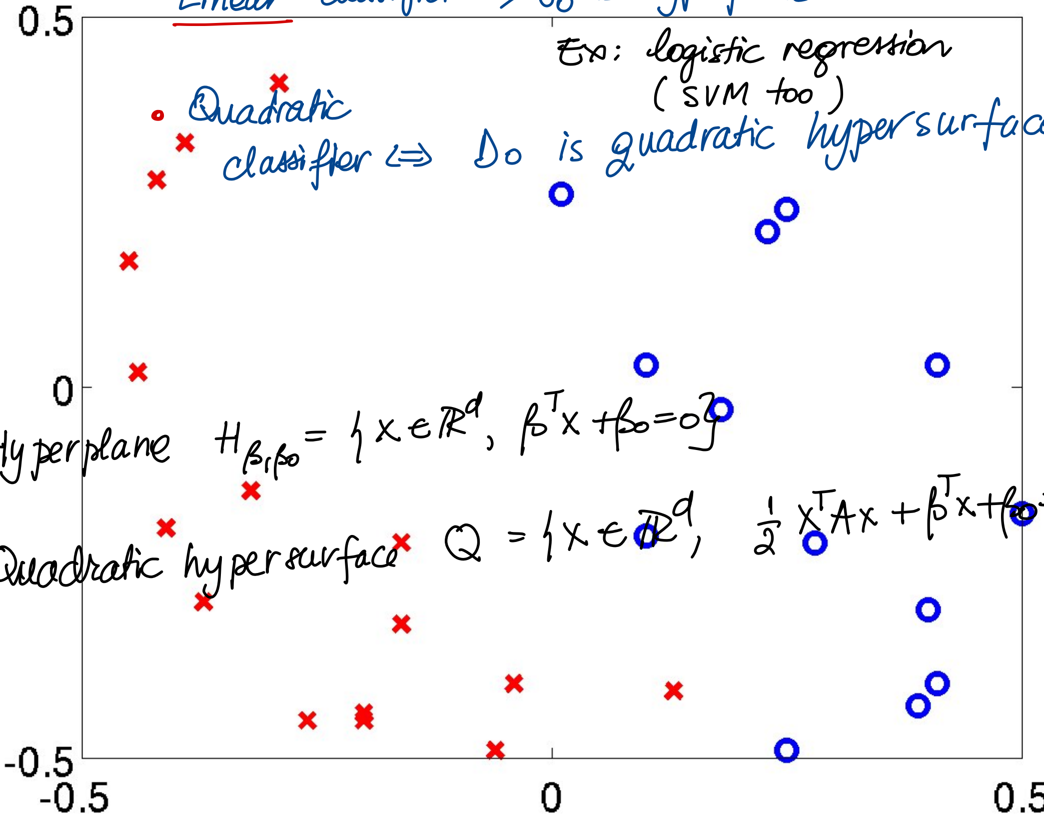


0.5
0
-0.5
-0.5
0
0.5
$-f(x) \Rightarrow D_0$ the same
$D_+ \leftarrow D_-$
$D_- \leftarrow D_+$
$f(x)<0$
$D_- \leftarrow$
$\rightarrow D_+$
$f(x)>0$
$D_0 = \{f(x)=0\}$
$f(x) = \beta^T x + \beta_0$
$\beta, x \in \mathbb{R}^2, \beta_0 \in \mathbb{R}$
$f_1(x) = 2\beta^T x + 2\beta_0$
same D_+, D_-, D_0
$f_2(x) = \text{logistic}(f(x)) - \frac{1}{2}$
same D_+, D_-, D_0

- **Linear** classifier ⇔ D_0 is hyperplane in $\mathbb{R}^d$ "linear"

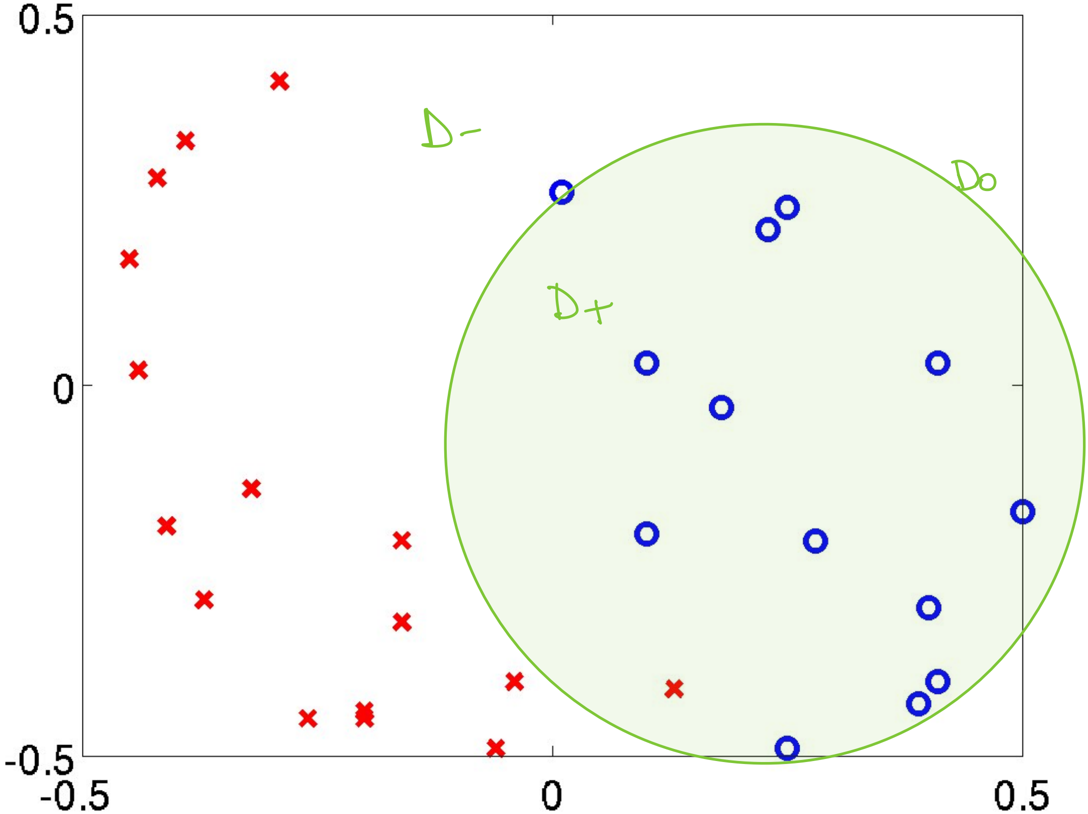
Ex: logistic regression (SVM too)

- Quadratic classifier ⇔ D_0 is quadratic hypersurface



Hyperplane $H_{\beta,\beta_0} = \{x \in \mathbb{R}^d,\ \beta^T x + \beta_0 = 0\}$

Quadratic hypersurface $Q = \{x \in \mathbb{R}^d,\ \frac{1}{2} x^T A x + \beta^T x + \beta_0 = 0\}$



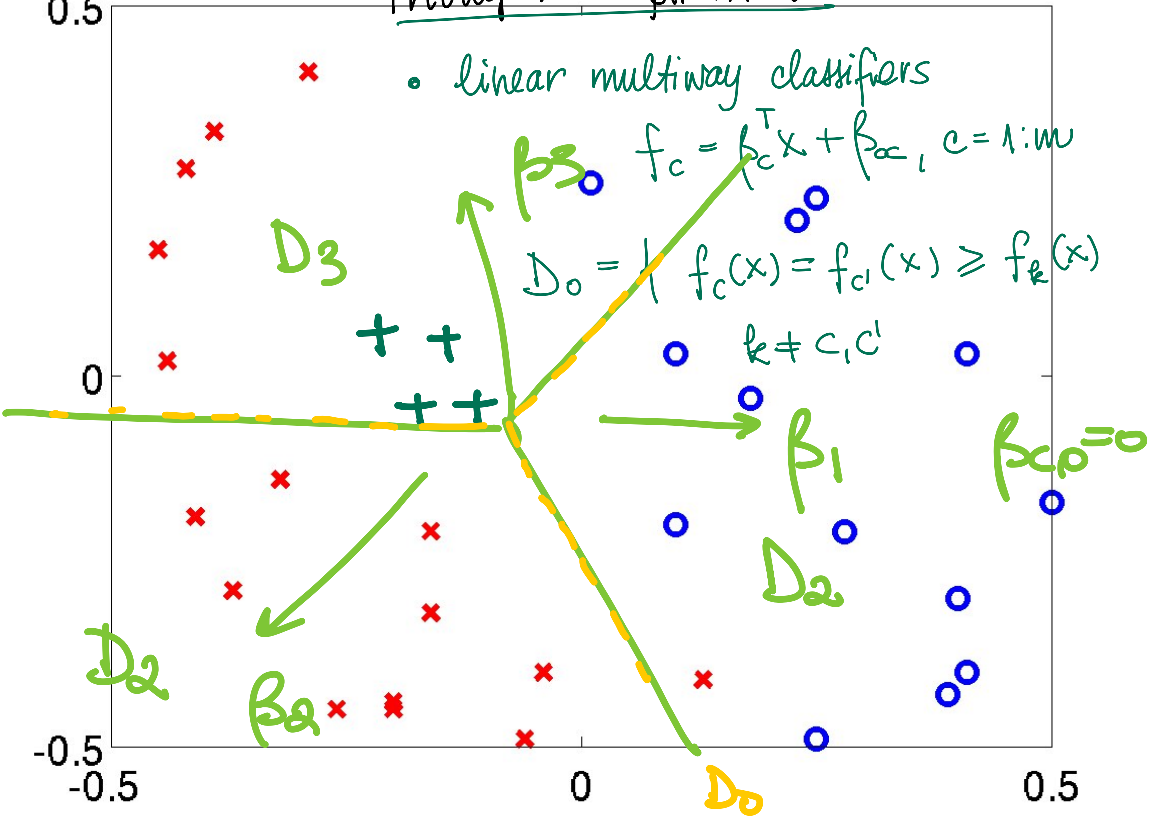
0.5
0
-0.5
-0.5
0
0.5
D-
Do
D+

Thought expriments

- linear multiway classifiers

$$f_c = \beta_c^T x + \beta_{0c}, \quad c = 1:m$$

$$D_0 = \{\, f_c(x) = f_{c'}(x) \geq f_k(x)$$
$$k \neq c, c'$$



Decision regions, decision boundary of a classifier

Let $f(x)$ be a classifier (not necessarily binary)

- $f(x)$ takes only a finite set of values
- The **decision region** associated with class y = the region in X space where f takes value y, i.e. $D_y = \{x \in \mathbb{R}^d,\ f(x) = y\} = f^{-1}(y)$.
- The boundaries separating the decision regions are called **decision boundaries**.

- For a binary classifier, we have two decision regions, D_+ and D_-. By convention $f(x) = 0$ on the decision boundary.
- For binary classifier with real valued $f(x)$ (i.e $\hat{y} = \mathrm{sgn} f(x)$) we define $D_+ = \{x \in \mathbb{R}^d,\ f(x) > 0\}$, $D_- = \{x \in \mathbb{R}^d,\ f(x) < 0\}$ and the decision boundary $\{x \in \mathbb{R}^d,\ f(x) = 0\}$

$$\ln\left(\underbrace{P[Y=1\,|\,x]\,/\,P[Y=-1\,|\,x]}\right) = \beta_0 + \beta_1 x + \frac{1}{2}\beta_2 x^2$$

$$D_0 = \{x,\ P[Y=1\,|\,x] = P[Y=-1\,|\,x]\}$$

$$\Leftrightarrow \beta_0 + \beta_1 x + \frac{1}{2}\beta_2 x^2 = 0$$ quadratic

dim $x > 1$ (parabola, hyperb, ellipse)