STAT 535

10/12/23

Lecture 5

- Q's on predictors
- Parametric vs Non-parametric
- Generative classifiers

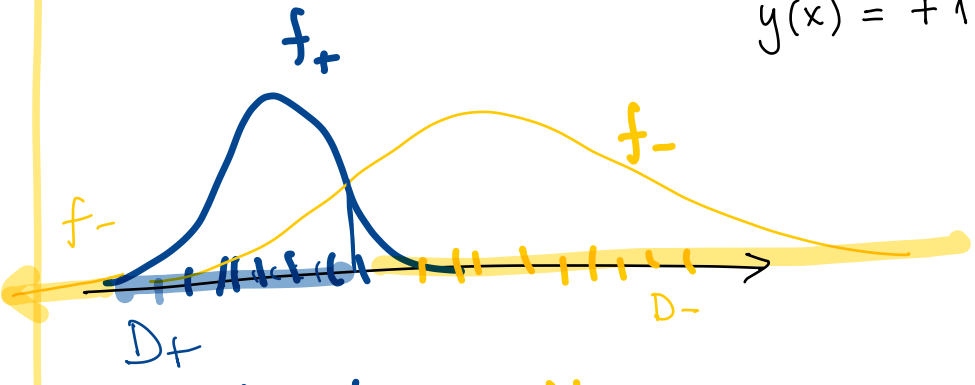
- Q1 on 10/19
- LII posted
- HW2 t.b. posted
  HW1 due [now, 10/19]

1. QDA intuition

2. LDA, QDA parameters

3. Decision Region DT

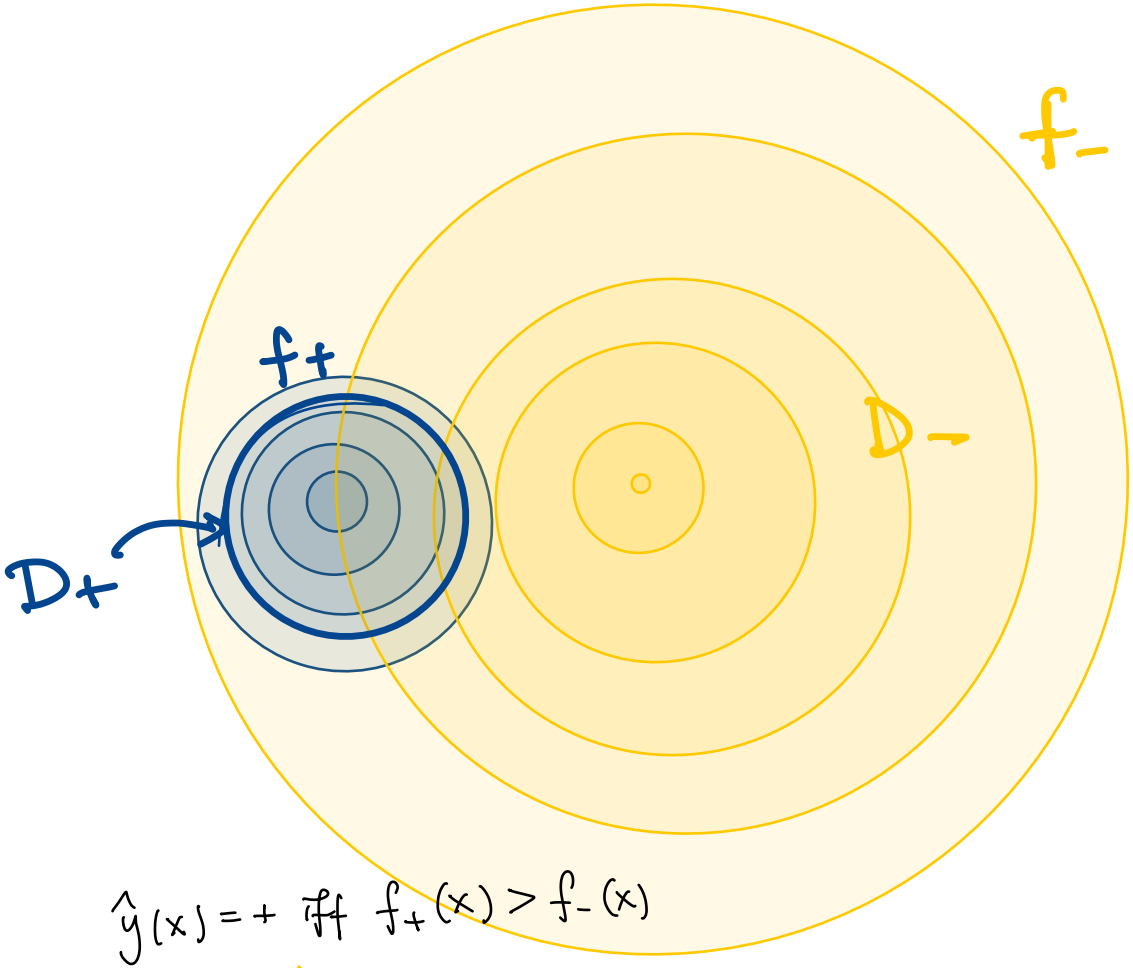
$\hat{y}(x) = +1$ iff $f_+(x) > f_-(x)$



$y = +$ $\quad$ $y = -$

$n_+ = n_- = \frac{n}{2}$

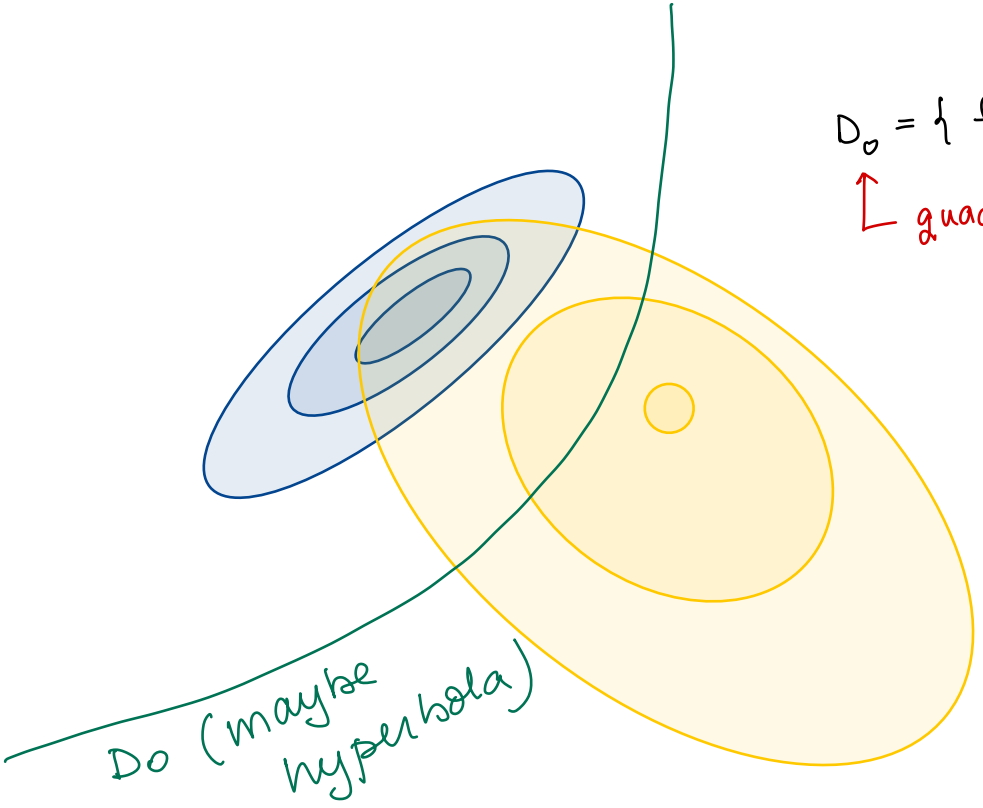
$\hat{\mu}_+, \hat{\Sigma}_+ = \hat{\sigma}_+^2$ $\quad$ $\hat{\mu}_-, \hat{\Sigma}_- = \hat{\sigma}_-^2$



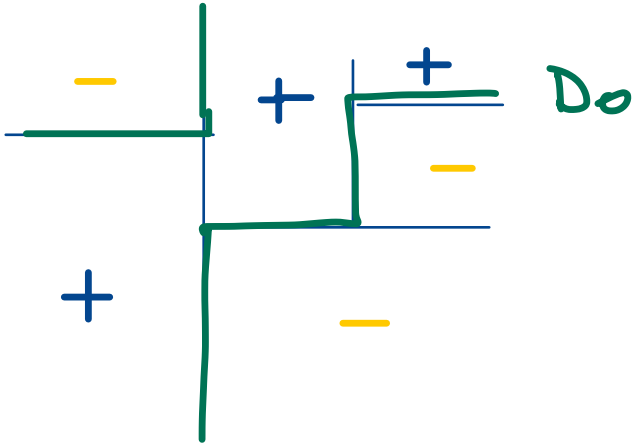
$$\hat{y}(x) = + \text{ if } f_+(x) > f_-(x)$$

$\hat{\Sigma}_- = \hat{\sigma}_-^2 I_p$

$\hat{\Sigma}_+ = \hat{\sigma}_+^2 I_p$


$D_0 = \{ f_+(x) = f_-(x) \}$
decision boundary
quadratic
D_0 (maybe hyperbola)

3. DT decision regions



Lecture II: Prediction – Basic concepts

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2023

Parametric vs non-parametric

Generative and discriminative models for classification

- Generative classifiers
- Discriminative classifiers
- Generative vs discriminative classifiers

Loss functions

- Bayes loss

Variance, bias and complexity

Reading HTF Ch.: 2.1–5,2.9, 7.1–4 bias-variance tradeoff, Murphy Ch.: 1., 8.6[1], Bach Ch.:

[1] Neither textbook is close to these notes except in a few places; take them as alternative perspectives or related reading

The "learning" problem

- **Given**
- a problem (e.g. recognize digits from $m \times m$ gray-scale images)
- a **sample** or (**training set**) of **labeled data**

$$\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$$

drawn i.i.d. from an unknown P_{XY}

- **model class** $\mathcal{F} = \{f\}$ = set of predictors to choose from

- **Wanted**
- a predictor $f \in \mathcal{F}$ that performs well on future samples from the same P_{XY}

  - "choose a predictor $f \in \mathcal{F}$" = training/learning
  - "performs well on future samples" (i.e. f **generalizes** well) – how do we measure this? how can we "guarantee" it?
  - choosing $\mathcal{F}$ is the **model selection problem** – about this later

A zoo of predictors

- Linear regression
- Logistic regression
- Linear Discriminant (LDA)
- Quadratic Discriminant (QDA)
- CART (Decision Trees)
- K-Nearest Neighbors
- Nadaraya-Watson (Kernel regression)
- Naive Bayes
- Neural networks/Deep learning
- Support Vector Machines
- Monotonic Regression

Classes of predictors
- by output ✓ { classification, regression
- by dimension { finite P, ∞ NP
- by "decision bdary" { generative, discriminative

Parametric vs. non-parametric models

Example (Parametric and non-parametric predictors)

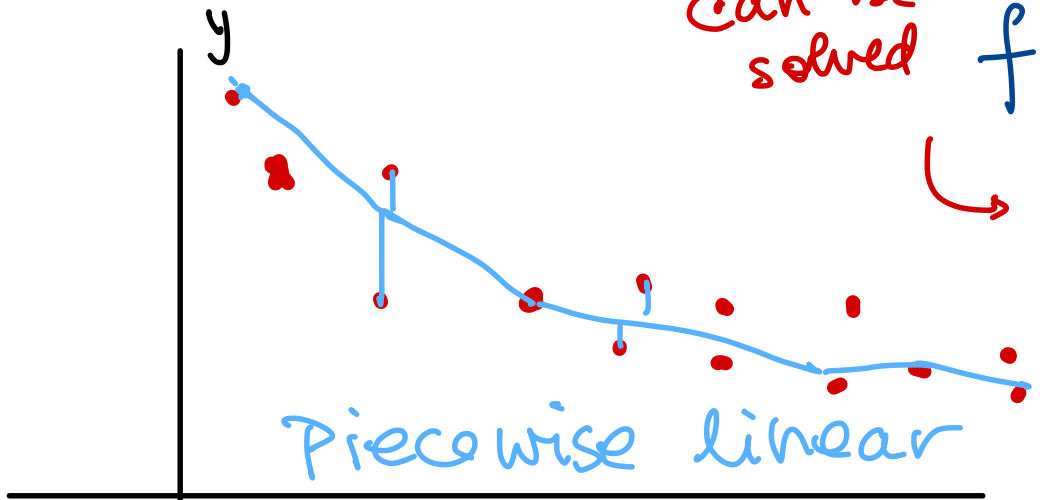
Parametric

- Linear, logistic regression ✓
- Linear Discriminant Analysis (LDA)
- Neural networks (small) ✓
- Naive Bayes ←
- CART with L levels = {Decision or leaves} {Regression tree

Non-parametric

- Nearest-neighbor classifiers and regressors ✓
- Nataraya-Watson predictors (kernel) ✓
- Monotonic regression ✓
- (Support Vector Machines)
- Large nn's

Exercise Are Radial Basis Functions classifiers parametric or non-parametric?



can be solved $\hat{f}(x) = \hat{y}$ monotonic

$\min_f \sum_{i=1}^{n} (y^i - f(x^i))^2$

f monotonic

"p" ∝ # segments ↗ with n, "noise"

$\mathcal{F} = \{ f \text{ decreasing} \}$

$\dim \mathcal{F} = \infty$

CART

L = # leaves = # splits + 1

- L fixed $\Rightarrow$ "p" fixed $\propto L$ $\Rightarrow$ $\mathcal{F}_L$ = {CART with L leaves} — parametric
- L grows with n $\Rightarrow$ $\mathcal{F}$ = {all CART} — non-parametric

e.g. $L = n$

Parametric
- (typically) interpretable
- [easier to analyse]
- convergence rate $n^{-1/2}$

Ex: LDA $\mu_\pm, \Sigma_\pm$

Ex: $\mathcal{D} \sim N(\mu, \sigma^2)$

wanted μ: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$E\left[\|\hat{\mu} - \mu\|^2\right] = \frac{\sigma^2}{n} \propto \frac{1}{n} = n^{-1}$$

$$\sqrt{\ldots} \propto n^{-1/2}$$

Parametric rate

Non-parametric
- (typically) not interpretable
- harder to analyse
- rate $n^{-\alpha}$ $\alpha < \frac{1}{2}$ slower

+ adapts to shape of data

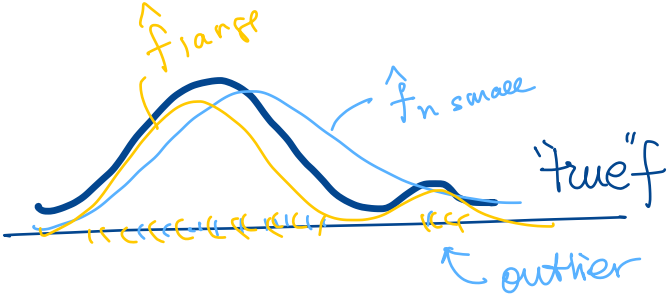
move formally later

How fast error $\searrow$ with n?

$\Updownarrow$

How large n for $E[\text{err}^2] = \varepsilon$?

if $\varepsilon' \leftarrow \varepsilon/10$ how large n'? $\Rightarrow \frac{n'}{n} = 10$



n small

n large

more formally later

$\hat{f} \to f$ for $n \to \infty$, smoothness $\searrow 0$

↳ optimal f "true"

Smoothness parameter

- chosen
- K-NN : K ↗ smoothness ↗
- kernel predictors : h ↗ smoothness ↗
- CART : L ↗ smoothness ↘

A mathematical definition

- A model class $\mathcal{F}$ is **parametric** if it is finite-dimensional, otherwise it is **non-parametric**

↳ $\mathcal{F}$ = vector space of f's

In other words

- When we estimate a parametric model from data, there is a fixed number of parameters, (you can think of them as one for each dimension, although this is not always true), that we need to estimate to obtain an estimate $\hat{f} \in \mathcal{F}$.
- The parameters are meaningful.
  E.g. the β_j in logistic regression has a precise meaning: the component of the normal to the decision boundary along coordinate j.
- The dimension of β does not change if the sample size n increases.

p = # parameters

Linear predictor

params $\beta_0, \beta \in \mathbb{R}^d \Rightarrow p = d+1$

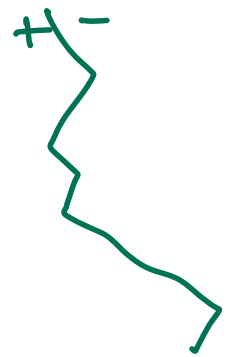
Parametric

independent of $n, \mathcal{D}$

Nonparametric

K-NN "p" = # params to describe $\mathcal{D}$

- grows with n
- growth depends on $\mathcal{D}$ (through P_{XY})



Non-parametric models – Some intuition

- When the model is non-parametric, the model class $\mathcal{F}$ is a function space.
- The $\hat{f}$ that we estimate will depend on some numerical values (and we could call them parameters), but these values have little meaning taken individually.
- The number of values needed to describe $\hat{f}$ generally grows with n.
  Examples In the Nearest neighbor and kernel predictors, we have to store all the data points, thus the number of values describing the predictor f grows (linearly) with the sample size. Exercise Does the number of values describing f always grow linearly with the sample size? Does it have to always grow to infinity? Does it have to always grow in the same way for a given $\mathcal{F}$?
- Non-parametric models often have a **smoothness parameter**.
  Examples of smoothness parameters K in K-nearest neighbor, h the kernel bandwidth in kernel regression.
  To make matters worse, a smoothness parameter is not a parameter! More precisely it is not a parameter of an $f \in \mathcal{F}$, because it is not estimated from the data, but a descriptor of the model class $\mathcal{F}$.
- We will return to smoothness parameters later in this lecture.

Generative classifiers

One way to define a classifier is to assume that each class is generated by a distribution $g_y(X) = P(X|Y = y)$. If we know the distributions g_y and the class probabilities $P(Y = y)$, we can derive the *posterior probability* distribution of Y for a given x. This is

$$P(Y = y|X) = \frac{P(Y = y)g_y(X)}{\sum_{y'} P(Y = y')g_{y'}(X)} = \frac{P(Y = y)g_y(X)}{P(X)} \quad (1)$$

The "best guess" for $Y(X)$ (i.e. the decision rule) is

$$f(X) = \text{argmax}_y P(Y = y|x) = \text{argmax}_y P(Y = y)g_y(x) \quad (2)$$

- (1) amounts to a likelihood ratio test for Y.
- The functions $g_y(x)$ are known as **generative models** for the classes y. Therefore, the resulting classifier is called a **generative classifier**. Examples: LDA, QDA, Naive Bayes.
- In contrast, a classifier defined directly in terms of $f(x)$ (or $P_{Y|X}$), like the linear, quadratic, decision tree is called a **discriminative classifier**.
- In practice, we may not know the functions $g_y(x)$, in which case we estimate them from the sample $\mathcal{D}$.

Generative classifiers (Binary)

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

1. fit $g_+ \approx P_{X|Y=+}$ ← generative models

   $g_- \approx P_{X|Y=-}$

   ↑ density or PMF

Ex: Classify objects by images $\quad g_{car} = f(\text{image} \mid \text{car})$

$g_{person} = f(\text{—"—} \mid \text{person})$

2. $p_\pm = \frac{n_\pm}{n}$ $\quad n_\pm = |\{x^i \mid y^i = \pm\}|$

   counts

3. Bayes' rule

$$P[y = + \mid x] = \frac{\overbrace{p_+}^{\text{prior}} \overbrace{g_+(x)}^{\text{likelihood}}}{p_+ g_+(x) + p_- g_-(x)}$$

Generative classifier and the likelihood ratio

$$P(Y = y|X) = \frac{P(Y = y)g_y(X)}{\sum_{y'} P(Y = y')g_{y'}(X)} = \frac{P(Y = y)g_y(X)}{P(X)}$$

$$f(x) = \text{argmax}_y P(Y = y|x) = \text{argmax}_y g_y(x)P(Y = y)$$

Likelihood Ratio test (for $y \in \{\pm 1\}$)

$$\frac{g_+(x)P(Y = +)}{g_-(x)P(Y = -)}$$

Example (Fisher's LDA in one dimension)

Assume $Y = \pm 1$, $g_y(x) = N(x, \pm\mu, \sigma^2 I)$, i.e each class is generated by a Normal distribution with the same spherical covariance matrix, but with a different mean. Let $P(Y = 1) = p \in (0, 1)$. Then, the posterior probability of Y is

$$P(Y = 1|x) \propto pe^{-||x-\mu||^2/(2\sigma^2)} \quad P(Y = -1|x) \propto (1-p)e^{-||x+\mu||^2/(2\sigma^2)} \tag{3}$$

and $f(x) = 1$ iff $\ln P(Y = 1|x)/P(Y = -1|x) \geq 0$, i.e iff

$$\ln \frac{p}{1-p} - \frac{1}{2\sigma^2}[||x^2|| - 2\mu^T x + ||\mu||^2 - ||x^2|| - (2\mu)^T x - ||\mu||^2] = \left(\frac{2\mu}{\sigma^2}\right)^T x + \ln \frac{p}{1-p} \geq 0 \tag{4}$$

Hence, the classifier $f(x)$ turns out to be a linear classifier. The decision boundary is perpendicular to the segment connecting the centers $\mu, -\mu$. This classifier is known as **Fisher's Linear Discriminant**. [**Exercises** Show that if the generative models are normal with different variances, then we obtain a quadratic classifier. What happens if the models g_y have the same variance, but it is a full covariance matrix Σ?]