STAT 391

2/19/25

Lecture 13

LV1 Model Sel
TB posted

l12-feb13 posted

Q2 2/27
≤ Model Selection

CV
Model Selection

Lecture Notes V – Non-parametric density estimation

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

January, 2025

Kernels

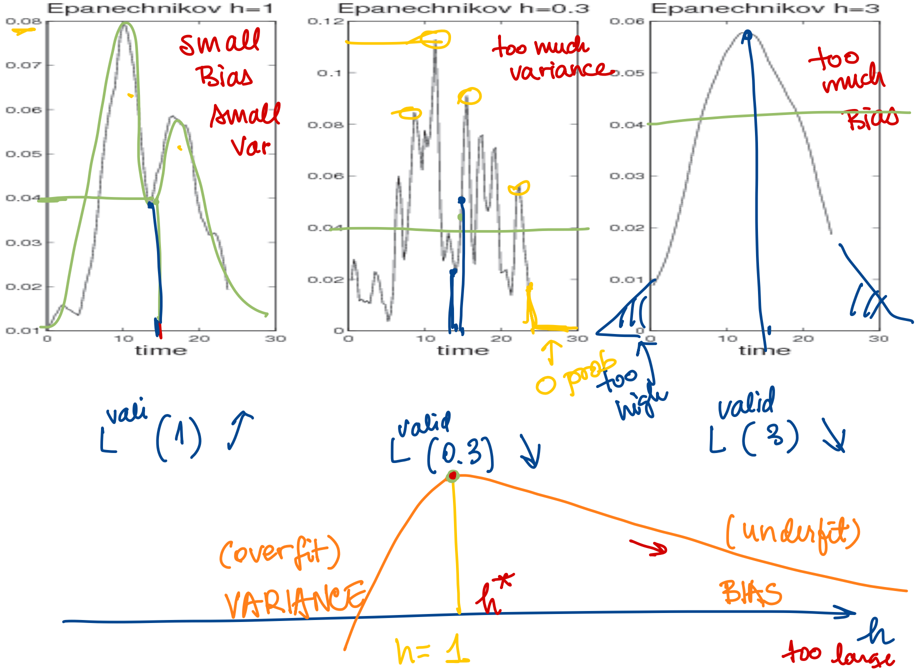
Kernel density estimators

Choosing h by Cross-Validation and the Bias-Variance trade-off

k-fold

Reading: Ch. 7

Choosing h



Epanechnikov h=1
Epanechnikov h=0.3
Epanechnikov h=3
time
Small Bias
Small Var
too much variance
too much Bias
0 prob
too high
vali L(1)
valid L(0.3)
valid L(3)
(overfit)
VARIANCE
(underfit)
BIAS
h*
h = 1
h too large

~~The Bias Variance trade-off~~

Choosing h in: $f_h = KDE(k(), h)$ for $\mathcal{D}$ given

model class

ideal: $f_h \approx f^{\text{unkn}}$

optimize

practical: sample $\tilde{\mathcal{D}} = \tilde{x}^1, \dots \tilde{x}^{n'} \sim f$

NEW dataset (new samples)

$$\max_h \text{Likelihood}(\underbrace{\tilde{x}^1 \dots \tilde{x}^{n'}}_{\text{validation data}} \mid f_h)$$

training data $x^{1:n}$

CROSS-VALIDATION

for h in some set

compute $L(\tilde{\mathcal{D}} \mid f_h(\mathcal{D})) = L^{\text{valid}}(h)$

validation | training

validation likelih

$$h^* = \operatorname{argmax}_h L^{\text{valid}}(h)$$

The Bias-Variance trade-off

Effect of h, n fixed

- h too small $\Rightarrow$ overfitting

  $\text{Var}\, f_h(x)$ large (for all x)

- h too large $\Rightarrow$ underfit

  $\text{Bias}\, f_h(x)$ large $\leftarrow$ can't converge to true $f(x)$ when $n \to \infty$

Effect of n, h fixed

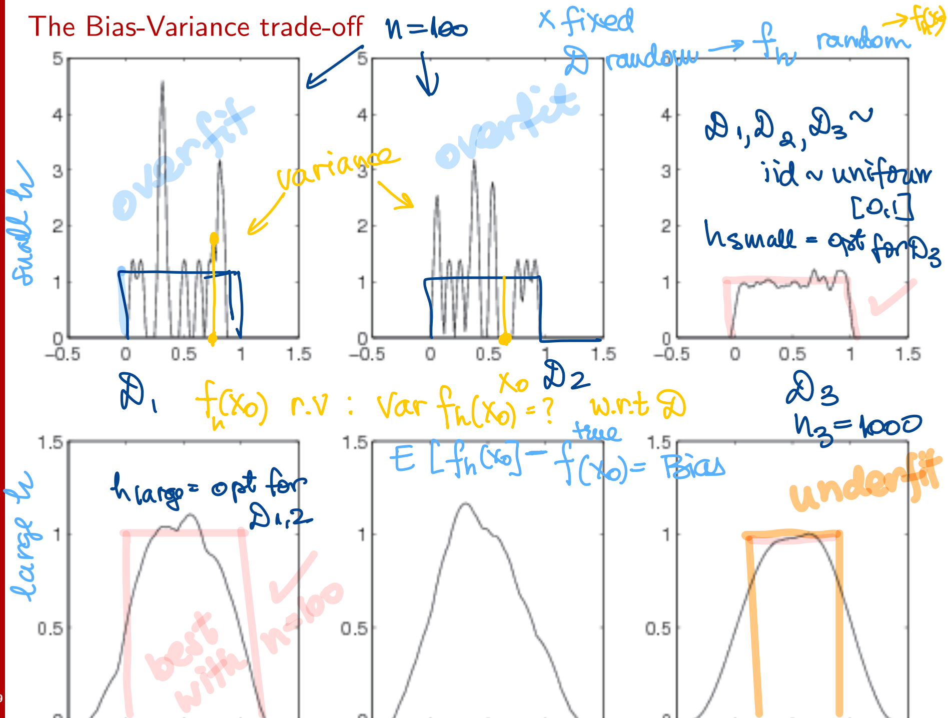
- $n \nearrow \Rightarrow \text{Var}\, f_h(x) \searrow$ for all h

  $\Downarrow$

  $h \searrow$ reduce bias

  $\Rightarrow$ a new balance btw Bias, Var

The Bias-Variance trade-off



K-fold Crossvalidation

In reality single $\mathcal{D}^{all} = \underbrace{\mathcal{D}}_{\text{train}} \overset{\text{split}}{\cup} \underbrace{\tilde{\mathcal{D}}}_{\text{validate}}$ $\qquad \mathcal{D} \cap \tilde{\mathcal{D}} = \emptyset$

$n^{all} = n + \tilde{n}$

random $\qquad$ random

ℓ^{cv} also random

n^{all} small:

K-fold CV

$K = 5, 10, n$

1. split $\mathcal{D}^{all}$ in K equal "folds"

$$\mathcal{D}^{all} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_K \qquad |\mathcal{D}_k| \cong \frac{n}{K}$$

model classes $\mathcal{F}_{1,2,3}$

2. for $k = 1, 2, \ldots K$

$$\tilde{\mathcal{D}} \leftarrow \mathcal{D}_k \quad \text{validate} \rightarrow \ell^{cv}(f_k^{1,2,3})$$

$$\mathcal{D} \leftarrow \mathcal{D}^{all} \setminus \mathcal{D}_k \quad \text{train} \rightarrow f_k^{1,2,3} \leftarrow n = \frac{K-1}{K} n^{all}$$

3. $$\ell^{cv}(f^{1,2,3}) = \frac{1}{K} \sum_{k=1}^{K} \ell^{cv}(f_k^{1,2,3})$$

final validation score for model class $\mathcal{F}$

Lecture Notes V – Model selection

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2021

k-fold

Cross-validation ✓

AIC and BIC

~~Structural risk minimization and VC dimension~~

~~**Reading** HTF Ch.. Ch. 7, Murphy Ch.. BIC, AIC 8.4.2 (pp 255), SRM 6.5 (pp 204)~~

Crossvalidation

What is Model selection

So far

Given $\mathcal{D} \sim$ iid on S from P_{unk}
Model class $\mathcal{F}$
$\Big\} \rightarrow$ estimate f

$\mathcal{F}_1 = \{\text{Normal}(\mu, \sigma^2)\}$ $\xrightarrow{ML}$ f_1

$\mathcal{F}_2 = \{\text{Logistic}\}$ $\xrightarrow{ML}$ f_2

$\mathcal{F}_3 = \{\text{mixture of gaussians}\}$ f_3

$\mathcal{F}_4 = \{\text{KDE}(k(), h)\}$ f_4

$\mathcal{F}_5 = \{\text{KDE}(k(), h')\}$ f_5

output

different model classes

• Which $\mathcal{F}$ to use?

Given

Model classes $\mathcal{F}_{1,2,3,\ldots}$

and $\left\{\begin{array}{l}\text{predictors} \\ \text{densities} \\ \text{density estimates}\end{array}\right.$ $f_1, f_2, \ldots$

[Optionally $\mathcal{D}$]

Wanted

which of $f_{1,2,\ldots}$ best predicts future data $\sim P_{unk}$

~~Crossvalidation~~

- Solution

1. Try all $\mathcal{F}_{1,2,\dots} \Rightarrow f_{1,2,\dots}$
2. Compare & select best of f_k's

$E_f[\ln f_k(x)]$ (*)

I → CV ← use $\tilde{D}$ in place of true f

II AIC, BIC ← cheap — no new $\tilde{D}$, no new computing — use math app

AIC → Akaike I.C (*)

BIC → Bayesian Information Criterion (**)

$E_{**}[\ln f_k(x)]$

more parameters, more overfit ← fewer params (smoother), more underfit

AIC and BIC

← only parametric models, with ML estimation (no KDE)

1. Fit models $f_1, f_2 \ldots$ (parametric)
2. compute $l(\underline{\mathcal{D}}|f_k) = l_k$

   TRAINING SET loglikelihood

3. Comparison

   how well f_k fits $\mathcal{D}$

   score $AIC(f_k) = l_k - d$ ← how many parameters used

   $BIC(f_k) = l_k - \frac{d \ln n}{2}$

   sample size

4. Choose $\underset{k}{\operatorname{argmax}} f$

$\boxed{d = \#\text{parameters}}$

$\mathcal{F}_1 = \{\text{normal}\} \rightarrow d_1 = 2$

$\mathcal{F}_2 = \{\text{mix } 2\text{ normal}\} \rightarrow d_2 = 2 \cdot 2 + 1$

$\mathcal{F}_3 = \{\text{mix } 3\text{ normal}\} \rightarrow d_3 = 2 \cdot 3 + 2$

$\mathcal{F}_4 = \{\text{logistic}\} \rightarrow d_4 = 2$

AIC and BIC

- $l_k = l_{k'} \Rightarrow$
  choose model
  with $d_k < d_{k'}$

- $d_k = d_{k'} \Rightarrow$
  choose model
  with $l_k > l_{k'}$

can happen
when e.g n very large