Statistical Modeling

Larry Wasserman
Department of Statistics and Data Science
and
Machine Learning Department
Carnegie Mellon University

Outline

Outline

1. Parametric Models

Outline

1. Parametric Models
2. Misspecified Parametric Models

Outline

1. Parametric Models
2. Misspecified Parametric Models
3. Nonparametric Models

Outline

1. Parametric Models
2. Misspecified Parametric Models
3. Nonparametric Models
4. Universal Inference (time permitting)

Parametric Models

Parametric Models

$$X_1, \ldots, X_n \sim p(x; \theta)$$

Parametric Models

$$X_1, \ldots, X_n \sim p(x; \theta)$$

Likelihood:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$$

Parametric Models

$$X_1, \ldots, X_n \sim p(x; \theta)$$

Likelihood:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$$

Maximum likelihood (MLE):

Parametric Models

$$X_1, \ldots, X_n \sim p(x; \theta)$$

Likelihood:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$$

Maximum likelihood (MLE):

Find $\widehat{\theta}$ to maximize $\mathcal{L}(\theta)$

Parametric Models

$$X_1, \ldots, X_n \sim p(x; \theta)$$

Likelihood:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$$

Maximum likelihood (MLE):
Find $\widehat{\theta}$ to maximize $\mathcal{L}(\theta)$

$$\widehat{\theta} \approx N(\theta, J)$$

where $J = I^{-1}$ and I is the Fisher information:

$$I_{jk} = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta)}{\partial \theta_j \partial \theta_k}\right].$$

Parametric Models

Parametric Models

Let

$$C = \left\{ \theta : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})} > e^{-(1/2)\chi^2_{d,\alpha}} \right\}.$$

Parametric Models

Let

$$C = \left\{ \theta : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})} > e^{-(1/2)\chi^2_{d,\alpha}} \right\}.$$

Then

$$P(\theta \in C) \to 1 - \alpha.$$

Parametric Models

Let

$$C = \left\{ \theta : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})} > e^{-(1/2)\chi^2_{d,\alpha}} \right\}.$$

Then

$$P(\theta \in C) \to 1 - \alpha.$$

For θ_j use $\widehat{\theta}_j \pm z_{\alpha/2}\sqrt{J_{jj}}$.

Parametric Models

Let

$$C = \left\{ \theta : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})} > e^{-(1/2)\chi^2_{d,\alpha}} \right\}.$$

Then

$$P(\theta \in C) \to 1 - \alpha.$$

For θ_j use $\widehat{\theta}_j \pm z_{\alpha/2}\sqrt{J_{jj}}$.

Or use the profile likelihood:

Parametric Models

Let

$$C = \left\{ \theta : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})} > e^{-(1/2)\chi^2_{d,\alpha}} \right\}.$$

Then

$$P(\theta \in C) \to 1 - \alpha.$$

For θ_j use $\widehat{\theta}_j \pm z_{\alpha/2}\sqrt{J_{jj}}$.

Or use the profile likelihood:

$$\mathcal{L}(\theta_j) = \sup_{\theta_{-j}} \mathcal{L}(\theta)$$

MLE: some useful facts

MLE: some useful facts

In general, the mle of $\psi(\theta)$ is $\psi(\widehat{\theta})$.

MLE: some useful facts

In general, the mle of $\psi(\theta)$ is $\psi(\widehat{\theta})$.

The MLE is asymptotically linear

MLE: some useful facts

In general, the mle of $\psi(\theta)$ is $\psi(\widehat{\theta})$.

The MLE is asymptotically linear

$$\sqrt{n}(\psi(\widehat{\theta}) - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_i \varphi(X_i) + o_P(1) \rightsquigarrow N(0, \mathbb{E}[\varphi(Z)\varphi(Z)^T])$$

where

$$\varphi(z) = \phi'(\theta) I_\theta^{-1} \ell'_\theta(x)$$

is the influence function and $\ell_\theta(x) = \log p_\theta(x)$.

MLE: some useful facts

MLE: some useful facts

The MLE is optimal:

MLE: some useful facts

The MLE is optimal:

if $\widetilde{\psi}$ is another (regular) estimator then

MLE: some useful facts

The MLE is optimal:

if $\widetilde{\psi}$ is another (regular) estimator then

$$\sqrt{n}(\widetilde{\psi} - \psi(\widehat{\theta})) \rightsquigarrow N(0, \mathbb{E}[\varphi(Z)\varphi(Z)^T]) + \text{noise}$$

MLE: some useful facts

The MLE is optimal:

if $\widetilde{\psi}$ is another (regular) estimator then

$$\sqrt{n}(\widetilde{\psi} - \psi(\widehat{\theta})) \rightsquigarrow N(0, \mathbb{E}[\varphi(Z)\varphi(Z)^T]) + \text{noise}$$

But this assumes:

MLE: some useful facts

The MLE is optimal:

if $\widetilde{\psi}$ is another (regular) estimator then

$$\sqrt{n}(\widetilde{\psi} - \psi(\widehat{\theta})) \rightsquigarrow N(0, \mathbb{E}[\varphi(Z)\varphi(Z)^T]) + \text{noise}$$

But this assumes:

(1) The model is correct.

MLE: some useful facts

The MLE is optimal:

if $\widetilde{\psi}$ is another (regular) estimator then

$$\sqrt{n}(\widetilde{\psi} - \psi(\widehat{\theta})) \rightsquigarrow N(0, \mathbb{E}[\varphi(Z)\varphi(Z)^T]) + \text{noise}$$

But this assumes:

(1) The model is correct.

(2) Regularity conditions.

Examples

Easy:

Examples

Easy:

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

Examples

Easy:

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

$\widehat{\mu} = \frac{1}{n} \sum_i X_i$

$\widehat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \overline{X})^2$

Examples

Easy:

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

$\widehat{\mu} = \frac{1}{n}\sum_i X_i$

$\widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$

Hard:

$$p(x) = (1-\lambda) \underbrace{b(x;\gamma)}_{\text{background}} + \lambda \underbrace{s(x;\theta)}_{\text{signal}}$$

$\widehat{\lambda}, \widehat{\gamma}, \widehat{\theta}$ must be found numerically

Problems: regularity conditions fail. More on this later.

Problems with the MLE

Problems with the MLE

Misspecified model

Problems with the MLE

Misspecified model

Not robust to outliers

Problems with the MLE

Misspecified model

Not robust to outliers

If there are many nuisance parameters, the MLE will be biased. (semiparametric models).

Problems with the MLE

Misspecified model

Not robust to outliers

If there are many nuisance parameters, the MLE will be biased. (semiparametric models).

High-dimensions: MLE no longer optimal.

Problems with the MLE

Misspecified model

Not robust to outliers

If there are many nuisance parameters, the MLE will be biased. (semiparametric models).

High-dimensions: MLE no longer optimal.

Failure of regularity conditions.

Misspecified Models

Misspecified Models

Suppose the model is wrong. (It usually is.) What are we estimating?

Misspecified Models

Suppose the model is wrong. (It usually is.) What are we estimating?

Define the Kullback-Leibler distance

$$K(p, q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx.$$

The mle converges to θ_* where θ_* minimizes $K(p_\theta, p)$.

Misspecified Models

Suppose the model is wrong. (It usually is.) What are we estimating?

Define the Kullback-Leibler distance

$$K(p, q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx.$$

The mle converges to θ_* where θ_* minimizes $K(p_\theta, p)$.

Is this a good thing?

Misspecified Models

Suppose the model is wrong. (It usually is.) What are we estimating?

Define the Kullback-Leibler distance

$$K(p,q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

The mle converges to θ_* where θ_* minimizes $K(p_\theta, p)$.

Is this a good thing?

Not necessarily.

Misspecified Models

Misspecified Models

Suppose use the Normal model

$$X_1, \ldots, X_n \sim N(\theta, 1).$$

Misspecified Models

Suppose use the Normal model

$$X_1, \ldots, X_n \sim N(\theta, 1).$$

Suppose the true distribution is

$$(1 - \epsilon)N(\theta, 1) + \epsilon\delta_a$$

where δ_a is some distribution concentrated at a point a.

Misspecified Models

Suppose use the Normal model

$$X_1, \ldots, X_n \sim N(\theta, 1).$$

Suppose the true distribution is

$$(1 - \epsilon)N(\theta, 1) + \epsilon\delta_a$$

where δ_a is some distribution concentrated at a point a.
Then $\theta_* \to \infty$ as $a \to \infty$.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Minimize

$$S(\theta) = \int p_0(x;\theta)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(X_i;\theta)^{\alpha}$$

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Minimize

$$S(\theta) = \int p_0(x;\theta)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(X_i;\theta)^{\alpha}$$

$\alpha \to 0$ gives mle.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Minimize

$$S(\theta) = \int p_0(x;\theta)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(X_i;\theta)^{\alpha}$$

$\alpha \to 0$ gives mle.

$\alpha > 0$ more robust.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Minimize

$$S(\theta) = \int p_0(x;\theta)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(X_i;\theta)^{\alpha}$$

$\alpha \to 0$ gives mle.

$\alpha > 0$ more robust.

$\alpha = 1$ is L_2: $\int (p - p_\theta)^2$.

Robustness to Model Misspecification

Maximum likelihood (Bayes) are not robust to model misspecification.

One solution: use power divergence (Basu, Harris, Hjort and Jones 1998):

Minimize

$$S(\theta) = \int p_0(x;\theta)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_i p_0(X_i;\theta)^{\alpha}$$

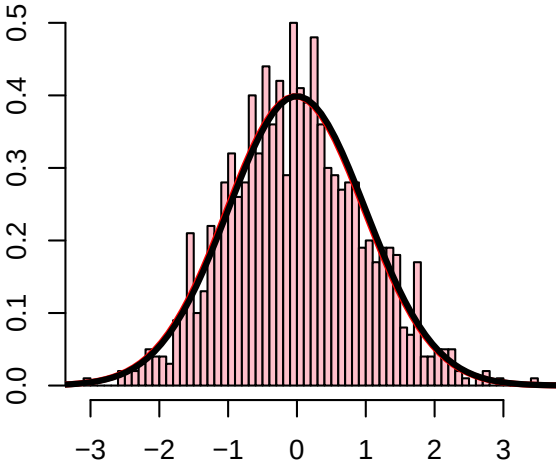
$\alpha \to 0$ gives mle.

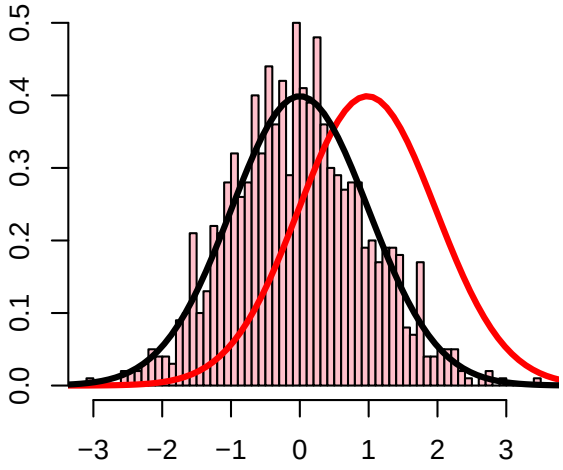
$\alpha > 0$ more robust.

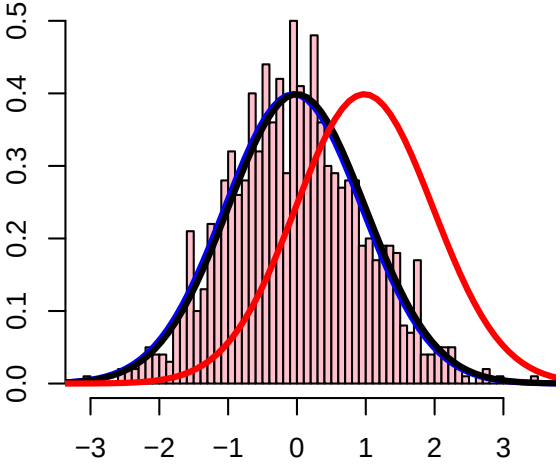
$\alpha = 1$ is L_2: $\int (p - p_\theta)^2$.

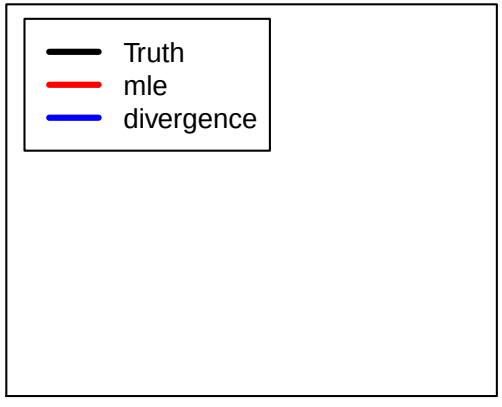
$\alpha \uparrow$: robustness $\uparrow$ and efficiency $\downarrow$

Example


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3


Truth
mle
divergence

Simulated Likelhoods

Suppose we want the likelihood ratio

$$R = \prod_{i=1}^{n} \frac{p(X_i; \theta_1)}{p(X_i; \theta_2)} = \prod_i R(X_i).$$

We may not have a closed form expression It is now common to estimate R by simulation + classification

Simulated Likelhoods

$$X_1, \ldots, X_N \sim p(x; \theta_1)$$
$$X_{N+1}, \ldots, X_{2N} \sim p(x; \theta_2)$$

Simulated Likelihoods

$$X_1, \ldots, X_N \sim p(x; \theta_1)$$
$$X_{N+1}, \ldots, X_{2N} \sim p(x; \theta_2)$$

S	1	1	$\cdots$	1	0	0	$\cdots$	0
X	X_1	X_2	$\cdots$	X_N	X_{N+1}	X_{N+2}	$\cdots$	X_{2N}

Simulated Likelhoods

Simulated Likelihoods

Classifier (neural net, random forest, ...)

$$h(x) = P(S = 1|X = x) = \frac{p(x;\theta_1)}{p(x;\theta_1) + p(x;\theta_2)} = \frac{1}{1 + R(x)}$$

Simulated Likelihoods

Classifier (neural net, random forest, ...)

$$h(x) = P(S = 1|X = x) = \frac{p(x;\theta_1)}{p(x;\theta_1) + p(x;\theta_2)} = \frac{1}{1 + R(x)}$$

$$R(x) = \frac{p(x;\theta_1)}{p(x;\theta_2)} = \frac{h(x)}{1 - h(x)}$$

Simulated Likelihoods

Classifier (neural net, random forest, ...)

$$h(x) = P(S = 1|X = x) = \frac{p(x;\theta_1)}{p(x;\theta_1) + p(x;\theta_2)} = \frac{1}{1 + R(x)}$$

$$R(x) = \frac{p(x;\theta_1)}{p(x;\theta_2)} = \frac{h(x)}{1 - h(x)}$$

$$\widehat{R} = \prod_{i=1}^{n} \frac{\widehat{h}(X_i)}{1 - \widehat{h}(X_i)}.$$

But, need $N > n^2$ to make sure that $\widehat{R} \approx R$.

Nonparametric Models: Density Estimation

$Y_1, \ldots, Y_n \sim p(y)$
$Y_i \in \mathbb{R}^d$

Nonparametric Models: Density Estimation

$Y_1, \ldots, Y_n \sim p(y)$
$Y_i \in \mathbb{R}^d$

We want to estimate the density $p(y)$ without assuming a model.

Nonparametric Models: Density Estimation

$Y_1, \ldots, Y_n \sim p(y)$
$Y_i \in \mathbb{R}^d$

We want to estimate the density $p(y)$ without assuming a model.

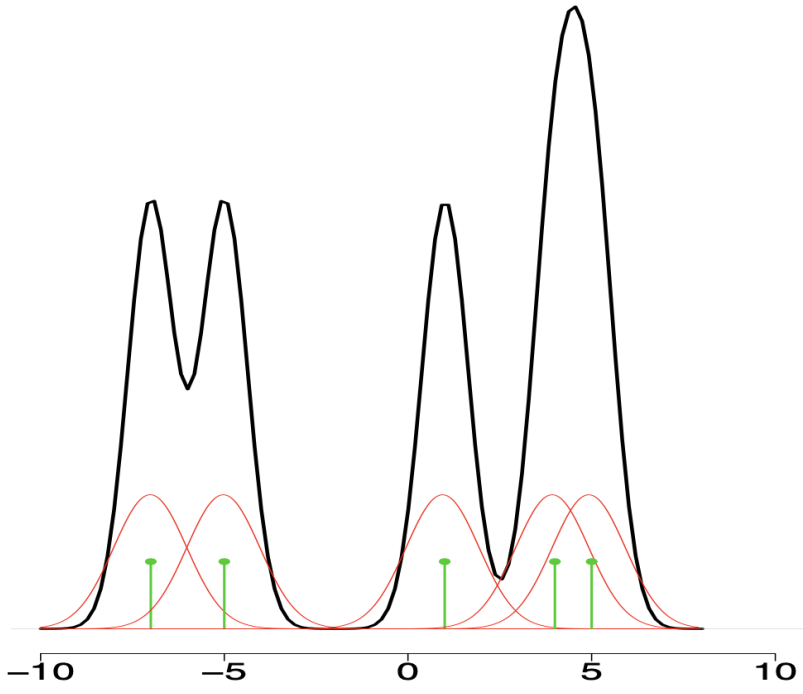
Kernel density estimator:

$$\widehat{p}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{Y_i - y}{h}\right)$$

where K is any kernel (example Gaussian) and $h > 0$ is the bandwidth.

Kernel Density Estimator

Places a smoothed-out lump of mass of size $1/n$ over each data point



The shape of the "lumps" is controlled by $K(\cdot)$; their width is controlled by h.

Density Estimation

MISE (mean integrated squared error)

$$\mathbb{E}[\int (\widehat{p}(x)-p(x))^2 dx] = \int \mathrm{bias}^2(x)dx + \int \mathrm{var}(x)dx = O(h^4)+O\left(\frac{1}{nh^d}\right)$$

where

$$\mathrm{bias}(x) = \mathbb{E}[\widehat{p}(x)] - p(x)$$

Density Estimation

MISE (mean integrated squared error)

$$\mathbb{E}[\int (\widehat{p}(x)-p(x))^2 dx] = \int \mathrm{bias}^2(x)dx + \int \mathrm{var}(x)dx = O(h^4)+O\left(\frac{1}{nh^d}\right)$$

where

$$\mathrm{bias}(x) = \mathbb{E}[\widehat{p}(x)] - p(x)$$

Best $h = n^{-2/(4+d)}$ gives
MISE $= O\left(\left(\frac{1}{n}\right)^{4/(4+d)}\right)$

Density Estimation

MISE (mean integrated squared error)

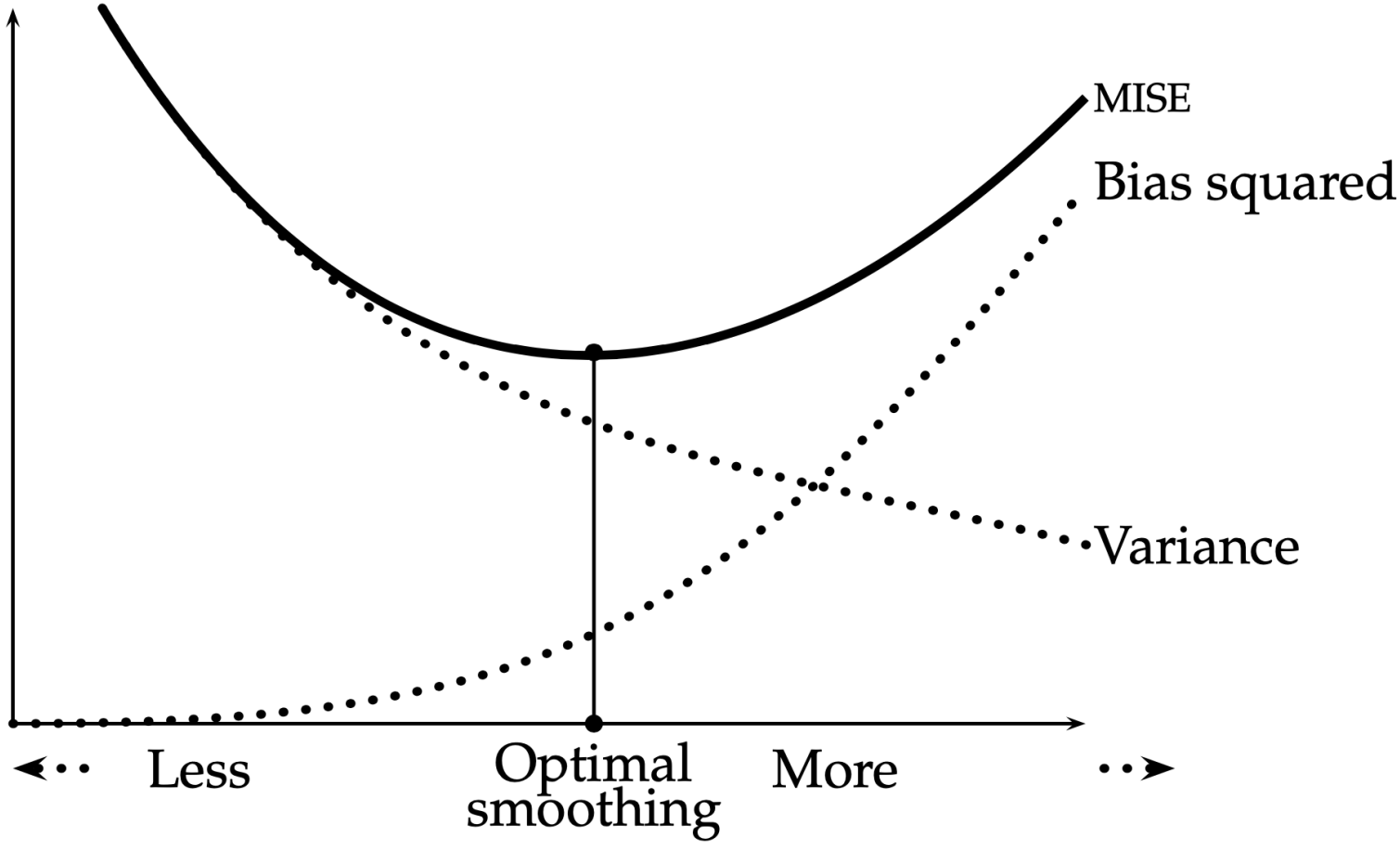
$$\mathbb{E}[\int (\widehat{p}(x)-p(x))^2 dx] = \int \mathrm{bias}^2(x)dx + \int \mathrm{var}(x)dx = O(h^4)+O\left(\frac{1}{nh^d}\right)$$

where

$$\mathrm{bias}(x) = \mathbb{E}[\widehat{p}(x)] - p(x)$$

Best $h = n^{-2/(4+d)}$ gives
MISE $= O\left(\left(\frac{1}{n}\right)^{4/(4+d)}\right)$
Faster rate than histograms.

Tradeoff



Choosing h by Cross Validation

Minimize

$$\int (\widehat{p}_h(x) - p(x))^2 dx = \int \widehat{p}_h^2(x) - 2 \int \widehat{p}_h(x) p(x) dx + C$$

Choosing h by Cross Validation

Minimize

$$\int (\widehat{p}_h(x) - p(x))^2 dx = \int \widehat{p}_h^2(x) - 2\int \widehat{p}_h(x)p(x)dx + C$$

Estimate by

$$\int \widehat{p}_h^2(x)dx - \frac{2}{n}\sum_i \widehat{p}_{h,(-i)}(X_i).$$

Nonparametric Regression (also called machine learning)

$(X_1, Y_1), \ldots, (X_n, Y_n)$
If Y is discrete, we call it classification (or machine learning).

Nonparametric Regression (also called machine learning)

$(X_1, Y_1), \ldots, (X_n, Y_n)$

If Y is discrete, we call it classification (or machine learning).

$\mu(x) = \mathbb{E}[Y|X = x]$.

Nonparametric Regression (also called machine learning)

$(X_1, Y_1), \ldots, (X_n, Y_n)$

If Y is discrete, we call it classification (or machine learning).

$\mu(x) = \mathbb{E}[Y|X = x]$.

Common estimators:

1. kernel
2. local polynomials
3. trees
4. random forests
5. neural nets/deep learning

Nonparametric Regression

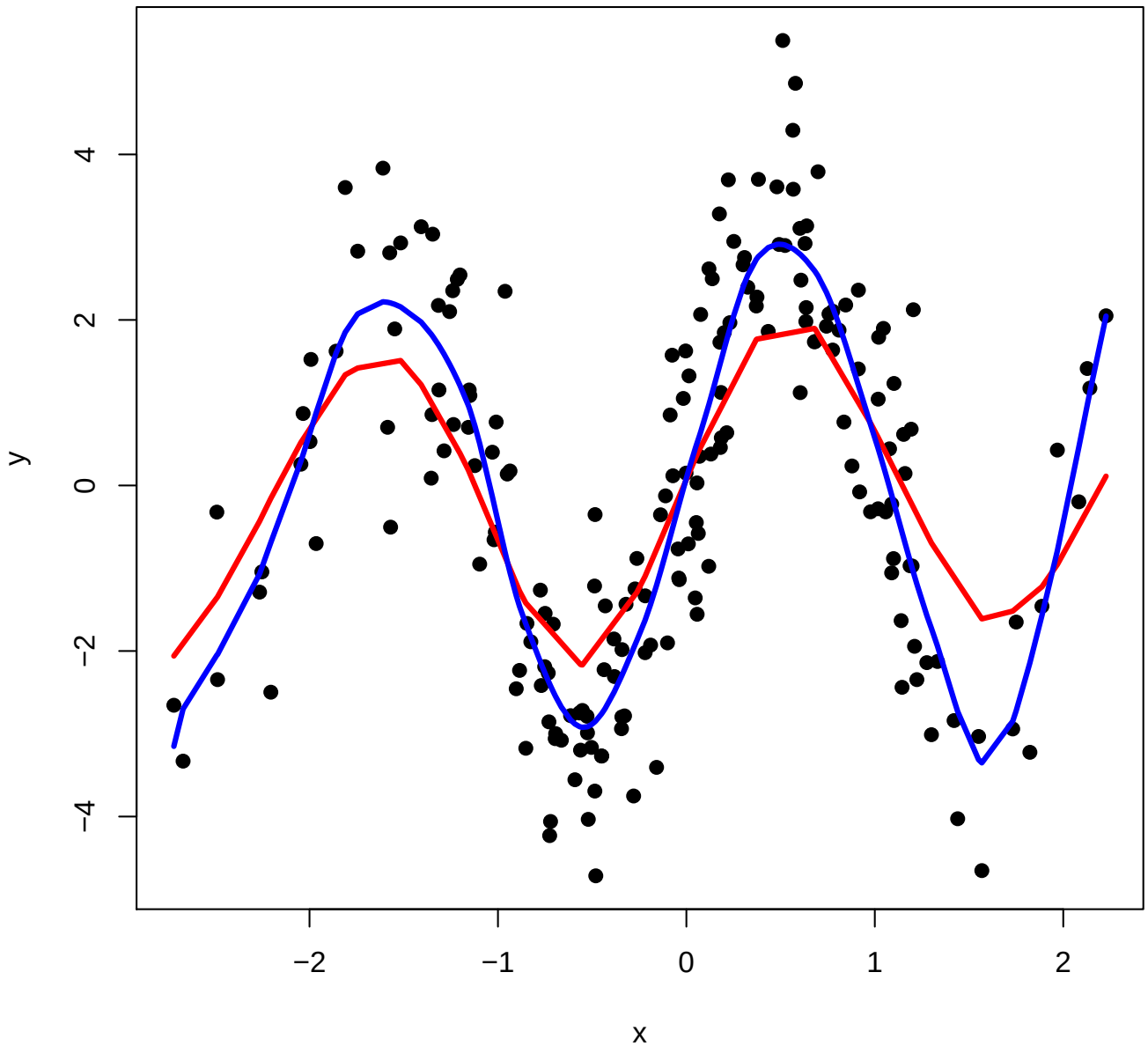
Kernel estimator

$$\widehat{\mu}(x) = \frac{\sum_i Y_i K_h(x - X_i)}{\sum_i K_h(x - X_i)} = \sum_i Y_i W_i(x)$$

Local polynomial:
$\mu(u) \approx \beta_0(x)(u - x) + \beta_1(x)(u - x)^2 + \cdots + \beta_p(x)(u - x)^p$
minimize

$$\sum_i (Y_i - [\beta_0(x)(X_i - x) + \cdots + \beta_p(x)(X_i - x)^p])^2 K_h(x - X_i)$$

Example



Trees

Construct an estimator $\widehat{\mu}(x)$ that is a piecewise constant of rectangles: $R_1, \ldots, R_N$ so that

$$\widehat{\mu}(x) = \sum_j \overline{Y}_j \, I(x \in R_j).$$

Trees

Construct an estimator $\widehat{\mu}(x)$ that is a piecewise constant of rectangles: $R_1, \ldots, R_N$ so that

$$\widehat{\mu}(x) = \sum_j \overline{Y}_j \, I(x \in R_j).$$

The rectangles are constructed recursively:
1. Find best split (which feature? where?)

Trees

Construct an estimator $\widehat{\mu}(x)$ that is a piecewise constant of rectangles: $R_1, \ldots, R_N$ so that

$$\widehat{\mu}(x) = \sum_j \overline{Y}_j \, I(x \in R_j).$$

The rectangles are constructed recursively:
1. Find best split (which feature? where?)

For example: $X_7 < 12.2$ versus $X_7 > 12.2$

Trees

Construct an estimator $\widehat{\mu}(x)$ that is a piecewise constant of rectangles: $R_1, \ldots, R_N$ so that

$$\widehat{\mu}(x) = \sum_j \overline{Y}_j \, I(x \in R_j).$$

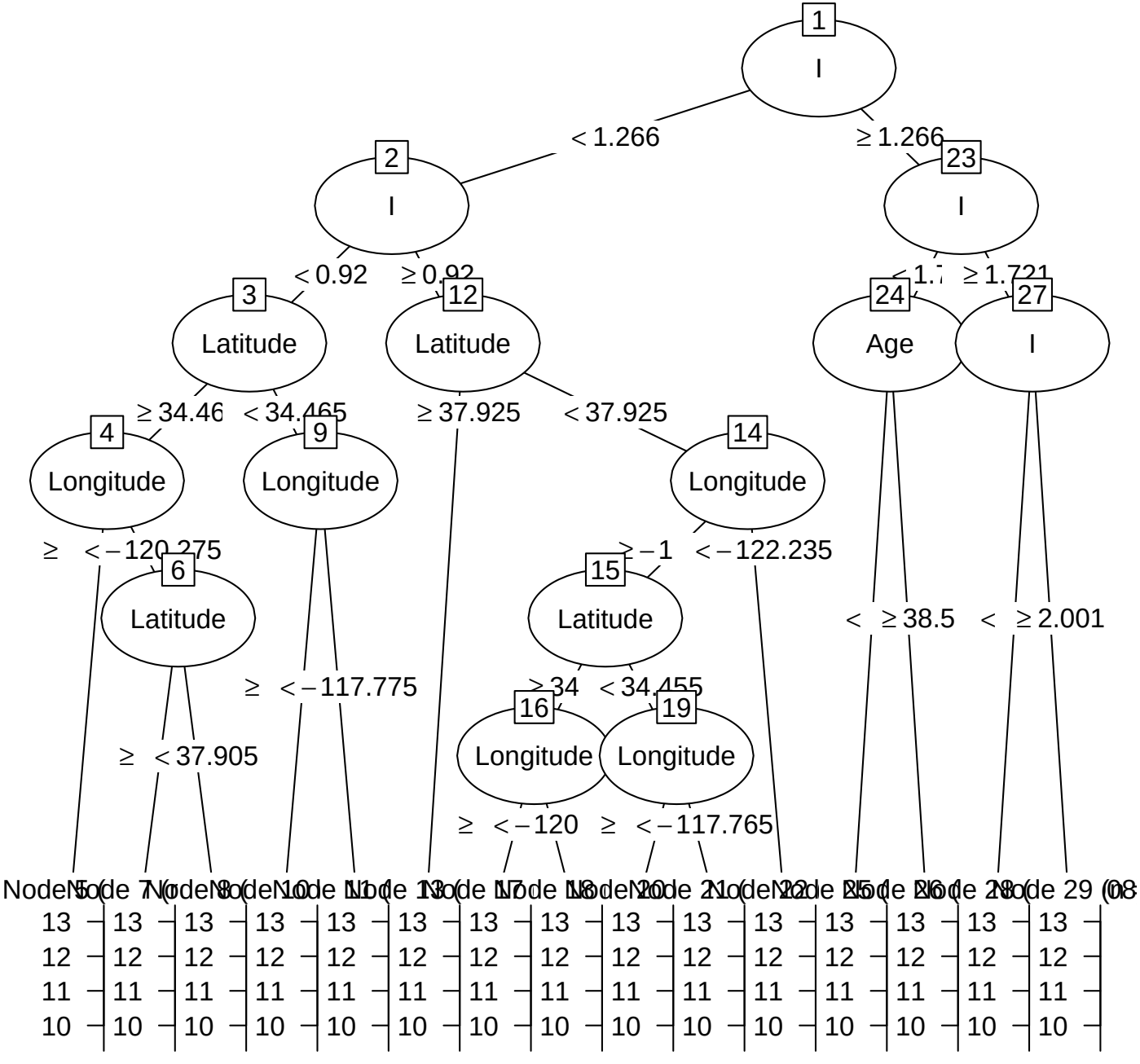
The rectangles are constructed recursively:

1. Find best split (which feature? where?)

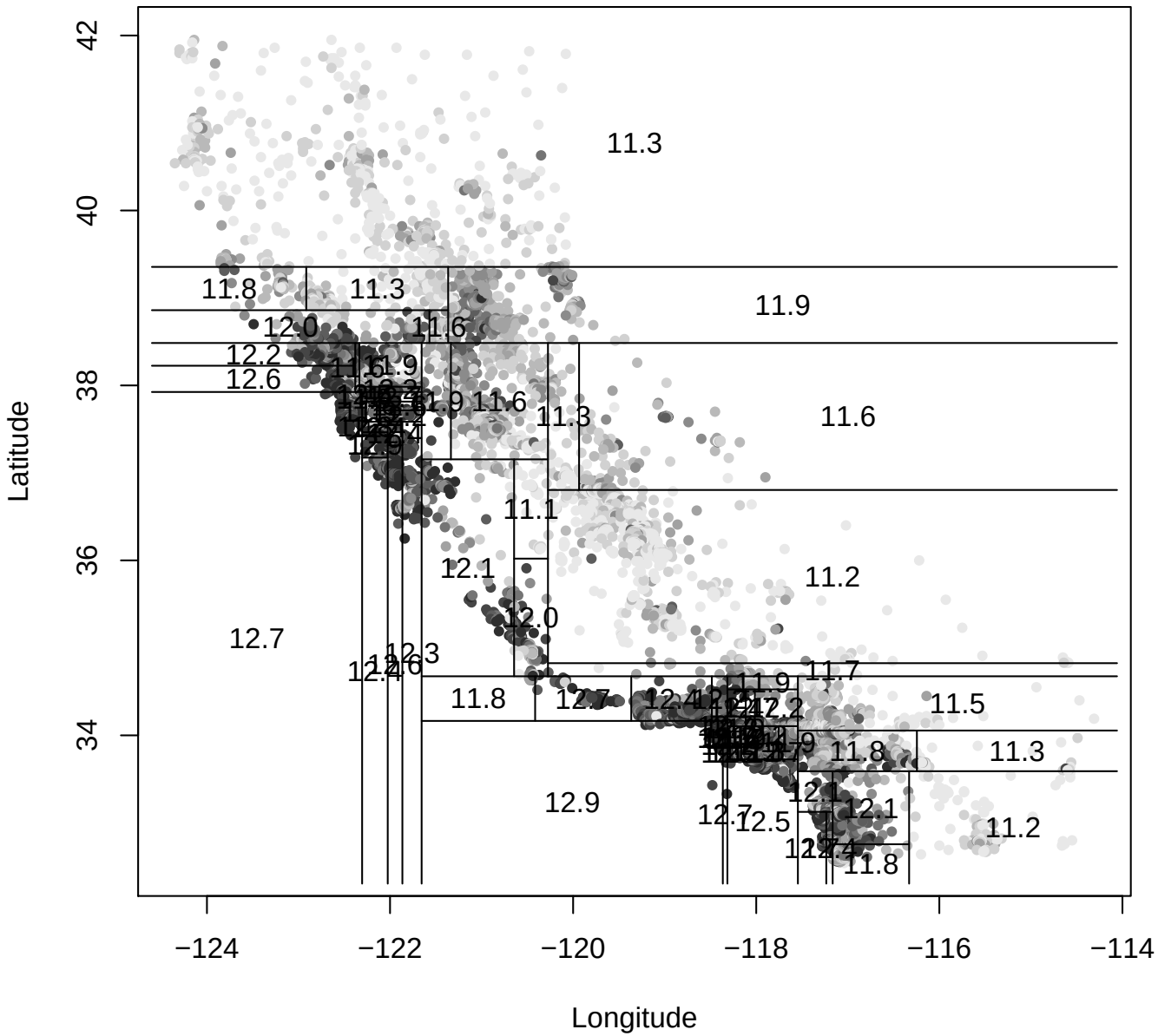
   For example: $X_7 < 12.2$ versus $X_7 > 12.2$

2. Repeat recursively.

Trees


1
I
< 1.266
≥ 1.266
2
I
23
I
< 0.92
≥ 0.92
≥ 1.721
3
Latitude
12
Latitude
24
Age
27
I
≥ 34.46
< 34.465
≥ 37.925
< 37.925
4
Longitude
9
Longitude
14
Longitude
≥ < −120.275
≥ −1
< −122.235
6
Latitude
15
Latitude
< ≥ 38.5
< ≥ 2.001
≥ < −117.775
≥ 34
< 34.455
≥ < 37.905
16
Longitude
19
Longitude
≥ < −120
≥ < −117.765
13
12
11
10

Trees


Latitude
Longitude
42
40
38
36
34
−124
−122
−120
−118
−116
−114
11.3
11.8
11.3
11.9
12.0
11.6
12.2
12.6
11.6
11.3
11.6
11.1
12.1
11.2
12.0
12.7
12.3
11.7
11.8
11.5
11.3
12.9
12.5
12.1
11.2
11.8

Forests: The Best Off-The-Shelf Method

Subsample data and features:

Forests: The Best Off-The-Shelf Method

Subsample data and features:

$S_1, \ldots, S_N$

Forests: The Best Off-The-Shelf Method

Subsample data and features:

$S_1, \ldots, S_N$

Each subsample S_j is obtained as follows:
randomly choose a subset of the features (for example, $d/3$ features):
then draw bootstrap sample of the data:
$(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$

Forests: The Best Off-The-Shelf Method

Subsample data and features:

$S_1, \ldots, S_N$

Each subsample S_j is obtained as follows:
randomly choose a subset of the features (for example, $d/3$ features):
then draw bootstrap sample of the data:
$(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$

Fit a tree on each subsample:

$$\widehat{\mu}_1(x), \ldots, \widehat{\mu}_N(x)$$

Forests: The Best Off-The-Shelf Method

Subsample data and features:

$S_1, \ldots, S_N$

Each subsample S_j is obtained as follows:
randomly choose a subset of the features (for example, $d/3$ features):
then draw bootstrap sample of the data:
$(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
Fit a tree on each subsample:

$$\widehat{\mu}_1(x), \ldots, \widehat{\mu}_N(x)$$

Then

$$\widehat{\mu}(x) = \frac{1}{N} \sum_{j=1}^{N} \widehat{\mu}_j(x)$$

A Word on Neural Nets/Deep Learning

A nonlinear regression model with many parameters.
For example:

$$\mu(x) = \alpha + \sum_j \sigma(\beta_j + \gamma_j^T X)$$

A Word on Neural Nets/Deep Learning

A nonlinear regression model with many parameters.
For example:

$$\mu(x) = \alpha + \sum_j \sigma(\beta_j + \gamma_j^T X)$$

Requires a lot of tuning and needs a lot of data

A Word on Neural Nets/Deep Learning

A nonlinear regression model with many parameters.
For example:

$$\mu(x) = \alpha + \sum_j \sigma(\beta_j + \gamma_j^T X)$$

Requires a lot of tuning and needs a lot of data

Try simpler methods first! In particular, try random forest. Much easier and often works very well.

Universal Inference (Wasserman, Ramdas and Balakrishnan 2020)

Universal Inference (Wasserman, Ramdas and Balakrishnan 2020)

Exact inference for hard statistical problems (failure of regularity conditions, nuisance parameters, asymptotics, discrete parameters)

Universal Inference (Wasserman, Ramdas and Balakrishnan 2020)

Exact inference for hard statistical problems (failure of regularity conditions, nuisance parameters, asymptotics, discrete parameters)

Exact (non asymptotic) coverage. No approximations. No regularity assumptions.

Background

The usual statistical tests and confidence intervals rely on specific mathematical assumptions.

Background

The usual statistical tests and confidence intervals rely on specific mathematical assumptions.

Toy example: $Y_1, \ldots, Y_n \sim N(\mu, 1)$

Background

The usual statistical tests and confidence intervals rely on specific mathematical assumptions.

Toy example: $Y_1, \ldots, Y_n \sim N(\mu, 1)$

Let $\psi = \mu^2$ and $\widehat{\psi} = \overline{Y}_n^2$.

Background

The usual statistical tests and confidence intervals rely on specific mathematical assumptions.

Toy example: $Y_1, \ldots, Y_n \sim N(\mu, 1)$

Let $\psi = \mu^2$ and $\widehat{\psi} = \overline{Y}_n^2$.

$\psi = 0$ behaves as χ^2
$\psi \neq 0$ behaves as Normal
$\psi \approx 0$ behaves as a mixture.
Valid confidence interval?

Example: Mixtures

$$p(x) = (1 - \lambda)N(\mu_1, 1) + \lambda N(\mu_2, 1)$$

(background + signal)

Example: Mixtures

$$p(x) = (1 - \lambda)N(\mu_1, 1) + \lambda N(\mu_2, 1)$$

(background + signal)

Null (no signal) =

$$\{(\mu_1, \mu_2, \lambda) : \ \lambda = 0\} = \{(\mu_1, \mu_2, \lambda) : \ \mu_1 = \mu_2\}$$

Example: Mixtures

$$p(x) = (1 - \lambda)N(\mu_1, 1) + \lambda N(\mu_2, 1)$$

(background + signal)

Null (no signal) =

$$\{(\mu_1, \mu_2, \lambda) : \ \lambda = 0\} = \{(\mu_1, \mu_2, \lambda) : \ \mu_1 = \mu_2\}$$

And when $\lambda = 0$, the parameter μ_2 is not identified.

Example: Mixtures

$$p(x) = (1 - \lambda)N(\mu_1, 1) + \lambda N(\mu_2, 1)$$

(background + signal)

Null (no signal) =

$$\{(\mu_1, \mu_2, \lambda) : \ \lambda = 0\} = \{(\mu_1, \mu_2, \lambda) : \ \mu_1 = \mu_2\}$$

And when $\lambda = 0$, the parameter μ_2 is not identified.

All of our standard machinery fails.

Other Issues With the Usual Methods

Standard methods are asymptotic (only valid as sample size $\longrightarrow \infty$)

Other Issues With the Usual Methods

Standard methods are asymptotic (only valid as sample size $\longrightarrow \infty$)

Nuisance parameters.

Other Issues With the Usual Methods

Standard methods are asymptotic (only valid as sample size $\longrightarrow \infty$)

Nuisance parameters.

Bootstrap: still requires regularity conditions and asymptotics.

Other Issues With the Usual Methods

Standard methods are asymptotic (only valid as sample size $\longrightarrow \infty$)

Nuisance parameters.

Bootstrap: still requires regularity conditions and asymptotics.

Bayes: does not provide coverage/error guarantees.

Universal Inference

Want confidence set C_n such that

Universal Inference

Want confidence set C_n such that

$P(\theta \in C_n) \geq 1 - \alpha$ for all θ and all n. No regularity conditions. Works with nuisance parameters.

Universal Inference

Want confidence set C_n such that

$P(\theta \in C_n) \geq 1 - \alpha$ for all θ and all n. No regularity conditions. Works with nuisance parameters.

Test $H_0 : \theta \in \Theta_0$.
Want: P(reject) under H_0 to be $\leq \alpha$. for all θ and all n. No regularity conditions. Works with nuisance parameters.

Universal Inference

Model $(p_\theta : \ \theta \in \Theta)$.

Universal Inference

Model $(p_\theta : \ \theta \in \Theta)$.

Split the data into two parts: D_0, D_1. (We'll get rid the splitting later.)

Universal Inference

Model $(p_\theta : \ \theta \in \Theta)$.

Split the data into two parts: D_0, D_1. (We'll get rid the splitting later.)

Get any estimate $\widehat{\theta}$ from D_1.

Universal Inference

Model $(p_\theta : \ \theta \in \Theta)$.

Split the data into two parts: D_0, D_1. (We'll get rid the splitting later.)

Get any estimate $\widehat{\theta}$ from D_1.

From D_0 construct

$$C = \left\{\theta : \ T \geq \alpha\right\}$$

where $T = \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})}$ and $L(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$.

Universal Inference

Model $(p_\theta : \ \theta \in \Theta)$.

Split the data into two parts: D_0, D_1. (We'll get rid the splitting later.)

Get any estimate $\widehat{\theta}$ from D_1.

From D_0 construct

$$C = \left\{\theta : \ T \geq \alpha\right\}$$

where $T = \frac{\mathcal{L}(\theta)}{\mathcal{L}(\widehat{\theta})}$ and $L(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$.

C is universal: $P(\theta \in C_n) \geq 1 - \alpha$
for all θ and all n. No asymptotics. No regularity conditions.

Getting rid of splitting

Split. Get T_1.
Split. Get T_2.
Split. Get T_3.
$\vdots$
Let

$$T = \frac{1}{B}\sum_j T_j$$

and let

$$C = \left\{\theta : \ T \geq \alpha\right\}$$

then C is universal.

Nuisance Parameters

$\theta = (\psi, \lambda)$

$$C = \left\{ \psi : \ T \geq \alpha \right\}$$

where now

$$T = \frac{L(\psi)}{L(\widehat{\theta})}$$

where

$$L(\psi) = \sup_{\lambda} L(\psi, \lambda).$$

Nuisance Parameters

$\theta = (\psi, \lambda)$

$$C = \left\{ \psi : \ T \geq \alpha \right\}$$

where now

$$T = \frac{L(\psi)}{L(\widehat{\theta})}$$

where

$$L(\psi) = \sup_{\lambda} L(\psi, \lambda).$$

Still valid.

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

$H_0 : \mu = \mu_0$.

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

$H_0 : \mu = \mu_0$.

D_1: any estimate $\widehat{\theta}$.

D_0: mle $\widehat{\theta}_0$ under H_0.

$$T = \frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)}$$

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

$H_0 : \mu = \mu_0$.

D_1: any estimate $\widehat{\theta}$.

D_0: mle $\widehat{\theta}_0$ under H_0.

$$T = \frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)}$$

Reject if $T \geq 1/\alpha$.

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

$H_0 : \mu = \mu_0$.

D_1: any estimate $\widehat{\theta}$.

D_0: mle $\widehat{\theta}_0$ under H_0.

$$T = \frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)}$$

Reject if $T \geq 1/\alpha$.

$P(\textit{reject})$ under H_0 is $\leq \alpha$ for all θ and n, no conditions. Average over splits to get rid of the randomness.

Hypothesis Testing With Nuisance Parameters

Want to test $H_0 : \theta \in \Theta_0$. For example, $\theta = (\mu, \lambda)$.

$H_0 : \mu = \mu_0$.

D_1: any estimate $\widehat{\theta}$.

D_0: mle $\widehat{\theta}_0$ under H_0.

$$T = \frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)}$$

Reject if $T \geq 1/\alpha$.

$P(\textit{reject})$ under H_0 is $\leq \alpha$ for all θ and n, no conditions. Average over splits to get rid of the randomness.

T is a p-value

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

Example:

$$(1 - \lambda)\gamma(x) + \lambda f(x; \theta)$$

where $\gamma(x)$ is a curve (background).

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

Example:

$$(1 - \lambda)\gamma(x) + \lambda f(x; \theta)$$

where $\gamma(x)$ is a curve (background).

This is called semiparametric inference.

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

Example:

$$(1 - \lambda)\gamma(x) + \lambda f(x; \theta)$$

where $\gamma(x)$ is a curve (background).

This is called semiparametric inference.

Another example:

$Y = \beta X + f(Z) + \epsilon$

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

Example:

$$(1 - \lambda)\gamma(x) + \lambda f(x; \theta)$$

where $\gamma(x)$ is a curve (background).

This is called semiparametric inference.

Another example:

$Y = \beta X + f(Z) + \epsilon$

This requires a different set of tools.

A Word on Semiparametric Methods

Consider a model like this:

$$p(y; \theta, \gamma)$$

where θ is finite dimensional and γ is infinite dimensional.

We want estimate and confidence interval for θ.

Example:

$$(1 - \lambda)\gamma(x) + \lambda f(x; \theta)$$

where $\gamma(x)$ is a curve (background).

This is called semiparametric inference.

Another example:
$Y = \beta X + f(Z) + \epsilon$

This requires a different set of tools.

No time to discuss, but be aware: usual methods don't work!

Conclusion

Parametric $\longrightarrow$ Semiparametric $\longrightarrow$ Nonparametric

Conclusion

Parametric $\longrightarrow$ Semiparametric $\longrightarrow$ Nonparametric

Parametric: MLE, least squares, method of moments, universal
Nonparametric: kernels, trees, forests, deep learning
Semiparametric: other tools

Conclusion

Parametric $\longrightarrow$ Semiparametric $\longrightarrow$ Nonparametric

Parametric: MLE, least squares, method of moments, universal
Nonparametric: kernels, trees, forests, deep learning
Semiparametric: other tools

Estimating likelihoods using classifiers. Great idea but be careful.

Conclusion

Parametric $\longrightarrow$ Semiparametric $\longrightarrow$ Nonparametric

Parametric: MLE, least squares, method of moments, universal
Nonparametric: kernels, trees, forests, deep learning
Semiparametric: other tools

Estimating likelihoods using classifiers. Great idea but be careful.

References:

Wasserman, L. All of Statistics

van der Vaart, A. Asymptotic Statistics.

Freedman, Hastie, Tibshirani. Elements of Statistical Learning.