Bayesian computation project

Charles Dufour

EPFL

Spring semester 2019

Table of contents

Framework implementations and limits[1]

Optimization

- gradient descent
- linear search gd
- Wolfe cond gd
- Newton gd (slow)
- ~~Stochastic gd~~

[1] more information can be found in the repository

Framework implementations and limits[1]

Optimization

- gradient descent
- linear search gd
- Wolfe cond gd
- Newton gd (slow)
- ~~Stochastic gd~~

Approximation

- Laplace
- GVA (unstable)

[1] more information can be found in the repository

Framework implementations and limits[1]

Optimization

- gradient descent
- linear search gd
- Wolfe cond gd
- Newton gd (slow)
- ~~Stochastic gd~~

Approximation

- Laplace
- GVA (unstable)

Sampling

- MH random walk
- MALA
- IS, RS
- ~~Gibbs~~
- ~~MH within Gibbs~~

[1] more information can be found in the repository

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

Purpose

- predict exactly the revenue
- predict if revenue above mean

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

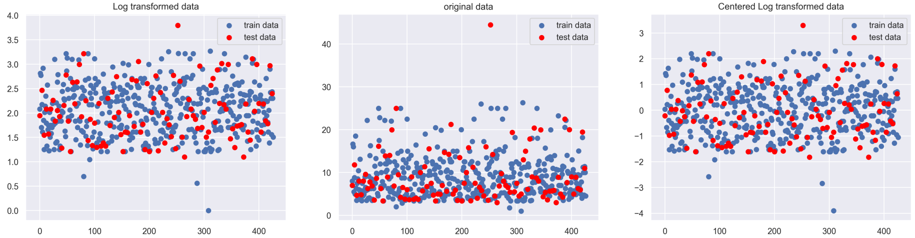
Purpose

- predict exactly the revenue
- predict if revenue above mean

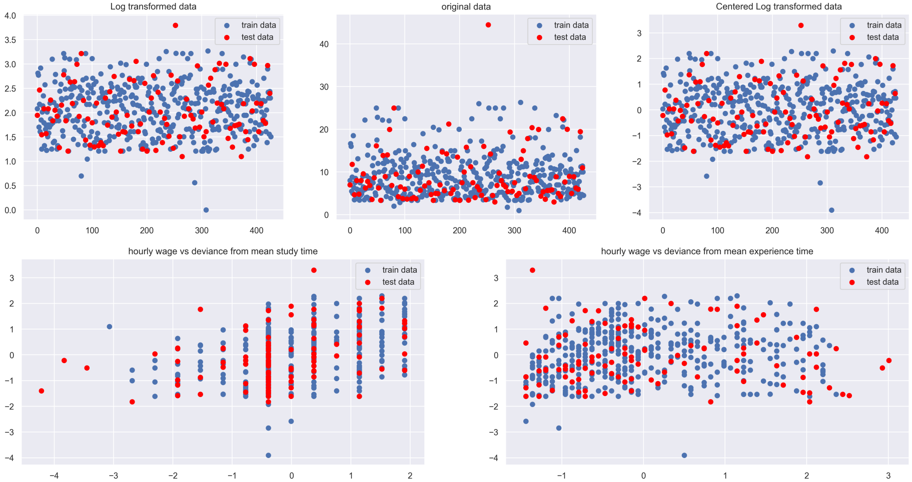
Features dropped due to high correlation

- AGE
- EXSQ

Visualization



Visualization



Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

- Student model

$$Y|\beta,\nu \sim X\beta + t_\nu \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \nu \sim \Gamma(2,4)$$

Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

- Student model

$$Y|\beta,\nu \sim X\beta + t_\nu \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \nu \sim \Gamma(2,4)$$

- Logistic regression

$$\mathbb{P}\left(Y=1|X,\beta\right) = \frac{e^{X^T\beta}}{1+e^{X^T\beta}}, \quad \beta \sim \mathcal{N}_d(0,3^2)$$

MAP (Maximum at posteriori) estimator

$$f(\theta) = -\log(\widetilde{f}(\theta|d))$$

$$\theta_c = \theta - \eta * \nabla_\theta f(\theta)$$

MAP (Maximum at posteriori) estimator

$$f(\theta) = -\log(\widetilde{f}(\theta|d))$$
$$\theta_c = \theta - \eta * \nabla_\theta f(\theta)$$

- Vanilla gradient descent

MAP (Maximum at posteriori) estimator

$$f(\theta) = -\log(\widetilde{f}(\theta|d))$$

$$\theta_c = \theta - \eta * \nabla_\theta f(\theta)$$

- Vanilla gradient descent
- Line search gd ($0 < \beta < 1$)
  accept if :
  $$f(\theta_c) < f(\theta) - \eta\varepsilon||\nabla_\theta f(\theta)||_2^2$$
  else:
  $$\eta = \beta\eta$$

MAP (Maximum at posteriori) estimator

$$f(\theta) = -\log(\widetilde{f}(\theta|d))$$

$$\theta_c = \theta - \eta * \nabla_\theta f(\theta)$$

- Vanilla gradient descent
- Line search gd ($0 < \beta < 1$)
  accept if :
  $$f(\theta_c) < f(\theta) - \eta\varepsilon||\nabla_\theta f(\theta)||_2^2$$
  else:
  $$\eta = \beta\eta$$
- Wolfe condition ($0 < c1 < c2 < 1$)
  decrease η if
  $$f(\theta_c) \geq f(\theta) + c_1\eta||\nabla_\theta f(\theta)||_2^2$$
  increase η if
  $$-\nabla_\theta f(\theta_c)^T\nabla_\theta f(\theta) \geq -c_2||\nabla_\theta f(\theta)||_2^2$$

MH with random walk

Theory

for $i = 1$ to N **do**
    draw $\eta \sim \mathcal{N}_d(0,1)$
    $\theta_c = \theta_n + \varepsilon\eta$
    $R = f(\theta_c|d)/f(\theta_n|d)$
    if $U(0,1) \leq R$ **then**
        $\theta_{n+1} = \theta_c$
    else
        $\theta_{n+1} = \theta_n$
    end if
end for

Practice

- set ε such that the acceptance rate of the proposal is between 10 and 50 percent.
- compute everything using `expsumlog`
- check visually the chain to determine the burn-in
- test different initialization to detect potential silent failure

MH with Langevin correction (MALA)

Theory

As for the random walk MH algorithm except for

- the proposal:

$$\theta_c = \theta_n + \tau \nabla \log f(\theta_n|d) + \sqrt{2\tau}\eta$$

- the acceptance ratio:

$$R = \frac{f(\theta_c|d)q(\theta_n|\theta_c)}{f(\theta_n|d)q(\theta_c|\theta_n)}$$

Practice

- as before but be more careful with tuning the step size τ
- $$q(x, x') \propto \exp\left(\frac{||x' - x - \tau\nabla \log f(x|d)||_2^2}{-4\tau}\right)$$
- biggest challenge: implement computation of gradient in efficient manner

Comparison of accuracy on test set

Gaussian MAE

	error on test	error on train
gd	0.685706	0.613765
line_search_gd	0.685706	0.613765
Wolfe_cond_gd	0.738525	0.692682
MH_vanilla_mean	0.687246	0.614395
MH_vanilla_median	0.688279	0.614266
MH_Langevin_mean	0.687873	0.613893
MH_Langevin_median	0.688071	0.614006

Comparison of accuracy on test set

Gaussian MAE

	error on test	error on train
gd	0.685706	0.613765
line_search_gd	0.685706	0.613765
Wolfe_cond_gd	0.738525	0.692682
MH_vanilla_mean	0.687246	0.614395
MH_vanilla_median	0.688279	0.614266
MH_Langevin_mean	0.687873	0.613893
MH_Langevin_median	0.688071	0.614006

Student MAE

	error on test	error on train
gd	0.685292	0.609907
line_search_gd	0.685159	0.611209
Wolfe_cond_gd	0.685284	0.609814
MH_vanilla_mean	0.687430	0.611630
MH_vanilla_median	0.685471	0.612177
MH_Langevin_mean	0.685110	0.611450
MH_Langevin_median	0.684638	0.611450

Comparison of accuracy on test set

Gaussian MAE

	error on test	error on train
gd	0.685706	0.613765
line_search_gd	0.685706	0.613765
Wolfe_cond_gd	0.738525	0.692682
MH_vanilla_mean	0.687246	0.614395
MH_vanilla_median	0.688279	0.614266
MH_Langevin_mean	0.687873	0.613893
MH_Langevin_median	0.688071	0.614006

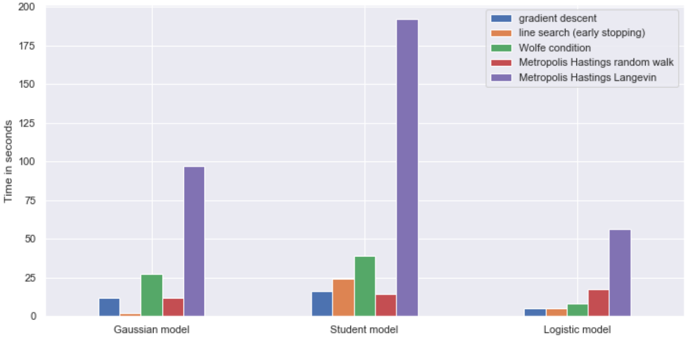
Student MAE

	error on test	error on train
gd	0.685292	0.609907
line_search_gd	0.685159	0.611209
Wolfe_cond_gd	0.685284	0.609814
MH_vanilla_mean	0.687430	0.611630
MH_vanilla_median	0.685471	0.612177
MH_Langevin_mean	0.685110	0.611450
MH_Langevin_median	0.684638	0.611450

Logistic error rate

	error on test	error on train
gd	0.448598	0.271663
line_search_gd	0.467290	0.269321
Wolfe_cond_gd	0.467290	0.269321
MH_vanilla_mean	0.467290	0.274005
MH_vanilla_median	0.467290	0.271663
MH_Langevin_mean	0.467290	0.271663
MH_Langevin_median	0.467290	0.274005

Comparison in term of time



Conclusion

Modelization

- simpler methods and models perform the best
- change the criterion : more realistic comparison

Conclusion

Modelization

- simpler methods and models perform the best
- change the criterion : more realistic comparison

Improvement

- Tuning of hyper-parameters
- Feature engineering

Conclusion

Modelization

- simpler methods and models perform the best
- change the criterion : more realistic comparison

Improvement

- Tuning of hyper-parameters
- Feature engineering
- Gamma model
- Classification in multiple ordered classes

Conclusion

Modelization

- simpler methods and models perform the best
- change the criterion : more realistic comparison

Improvement

- Tuning of hyper-parameters
- Feature engineering
- Gamma model
- Classification in multiple ordered classes
- More robust and faster module

References

- Guillaume Dehaene
  Lecture Notes, Bayesian computation MATH-435, 2019.
- E. R. Berndt
  The practice of econometrics : classic and contemporary. Addison-Wesley Pub. Co., 1991
- GitHub repository
  https://github.com/dufourc1/Bayesian computation