Chamberland-Dozois, Léandre (matricule : 1792798)

# INF8225 Rapport TP1

Travail présenté à M. Subramanian

Gr. 01

Polytechnique Montréal

4 février 2019

## 1.1

(1) Réseaux Bayésiens:

$$\Pr(A_1, A_2, \dots, A_N) = \prod_{i=1}^{N} \Pr\big(A_i \big| Parents(A_i)\big)$$

(2) Marginalisation :

$$Pr(X_1) = \sum_{\{X\}\setminus X_1} \Pr(X_1, X_2, \dots, X_N)$$

(3) Conditionnement :

$$\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$$

a)

$$\Pr(W=1) = \sum_{P} \Pr(W=1,P) \ \wedge \ \Pr(\mathrm{W}=1|\mathrm{P}) * \Pr(\mathrm{P}) \ = \ \Pr(\mathrm{W}=1,\mathrm{P}) \qquad \boldsymbol{eq.(2,3)}$$

$$\Rightarrow \ \Pr(W=1) = \sum_{P} \Pr(\mathrm{W}=1|\mathrm{P}) * \Pr(\mathrm{P})$$

L'équation de marginalisation est utilisée pour réduire la probabilité que le gazon de Watson à une sommation sur la probabilité conjointe de Watson et Pluie. Puisque la variable watson est dépendante conditionnellement de la variable pluie, on peut remplacer la probabilité conjointe par une probabilité conditionnelle connu Pr(W|P). On peut ensuite calculer la probabilité que le gazon de watson soit mouillé numériquement.

b)

$$\Pr(W=1|H=1) = \frac{\Pr(W=1,H=1)}{\Pr(H=1)} \ \wedge \qquad \boldsymbol{eq.(3)}$$

$$\Pr(H=1) = \sum_{A}\sum_{P}\sum_{W} \Pr(H=1,W,P,A) \ \wedge \qquad \boldsymbol{eq.(2)}$$

$$\Pr(W=1,H=1) = \sum_{A}\sum_{P} \Pr(H=1,W=1,P,A) \ \wedge \qquad \boldsymbol{eq.(2)}$$

$$\Pr(W,H,A,P) = \Pr(W|P) * Pr(P) * P(H|P,A) * P(A) \qquad \boldsymbol{eq.(1)}$$

$$\Rightarrow \ \Pr(W=1|H=1) = \frac{\sum_A \sum_P \Pr(W=1|P) * Pr(P) * P(H=1|P,A) * P(A)}{\sum_A \sum_P \sum_W \Pr(W|P) * Pr(P) * P(H=1|P,A) * P(A)}$$

Puisque Pr(W=1|H=1) n'est pas connu, on peut utiliser le conditionnement pour représenter la probabilité conditionnelle comme une division d'un probabilité conjointe sur la probabilité Pr(H=1). Le numérateur et le dénominateur peuvent être calculé à partir d'une marginalisation. Les deux marginalisations devraient contenir la probabilité conjointe complète du réseau pour pouvoir utiliser la définition. En utilisant la définition des réseau bayésiens, la probabilité Pr(W=1|H=1) peut être calculé à l'aide de sommations de multiplications de probabilités conditionnelles connues.

c)

$$\Pr(W=1|H=1,A=0)=\frac{\Pr(W=1,H=1,A=0)}{\Pr(H=1,A=0)} \wedge \quad \boldsymbol{eq}.(\mathbf{3})$$

$$\Pr(H=1,A=0)=\sum_P\sum_W \Pr(H=1,W,P,A=0) \wedge \quad \boldsymbol{eq}.(\mathbf{2})$$

$$\Pr(W=1,H=1,A=0)=\sum_P \Pr(H=1,W=1,P,A=0) \wedge \quad \boldsymbol{eq}.(\mathbf{2})$$

$$\Pr(W,H,A,P)=\Pr(W|P)*Pr(P)*P(H|P,A)*P(A) \quad \boldsymbol{eq}.(\mathbf{1})$$

$$\Rightarrow \Pr(W=1|H=1,A=0)=\frac{\sum_P \Pr(W=1|P)*Pr(P)*P(H=1|P,A=0)*P(A=0)}{\sum_P\sum_W \Pr(W|P)*Pr(P)*P(H=1|P,A=0)*P(A=0)}$$

Le même raisonnement que celui utilisé pour le b) s'applique ici. Par contre, on peut remarquer que l'observation H=1 crée une dépendance conditionnelle entre A et P. De plus, l'observation A=0 laisse penser que P=1 (concept «explaining away»).

d)

$$\Pr(W=1|\,A=0)=\frac{\Pr(W=1,\ A=0)}{\Pr(A=0)} \wedge \quad \boldsymbol{eq}.(\mathbf{3})$$

$$\Pr(W=1,A=0)=\sum_P\sum_H \Pr(H,W=1,P,A=0) \wedge \quad \boldsymbol{eq}.(\mathbf{2})$$

$$\Pr(A=0)=\sum_H\sum_W\sum_P \Pr(H,W,P,A=0) \wedge \quad \boldsymbol{eq}.(\mathbf{2})$$

$$\Pr(W,H,A,P)=\Pr(W|P)*Pr(P)*P(H|P,A)*P(A) \quad \boldsymbol{eq}.(\mathbf{1})$$

$$\Rightarrow \Pr(W=1|A=0)=\frac{\sum_P\sum_H \Pr(W=1|P)*Pr(P)*P(H|P,A=0)*P(A=0)}{\sum_H\sum_W\sum_P \Pr(W|P)*Pr(P)*P(H|P,A=0)*P(A=0)}$$

Le même raisonnement que pour le b) et le c) peut être utilisé ici. Par contre, puisqu'il n'y a pas d'observation sur H, on ne peut plus dire que A et P sont conditionnellement dépendants.

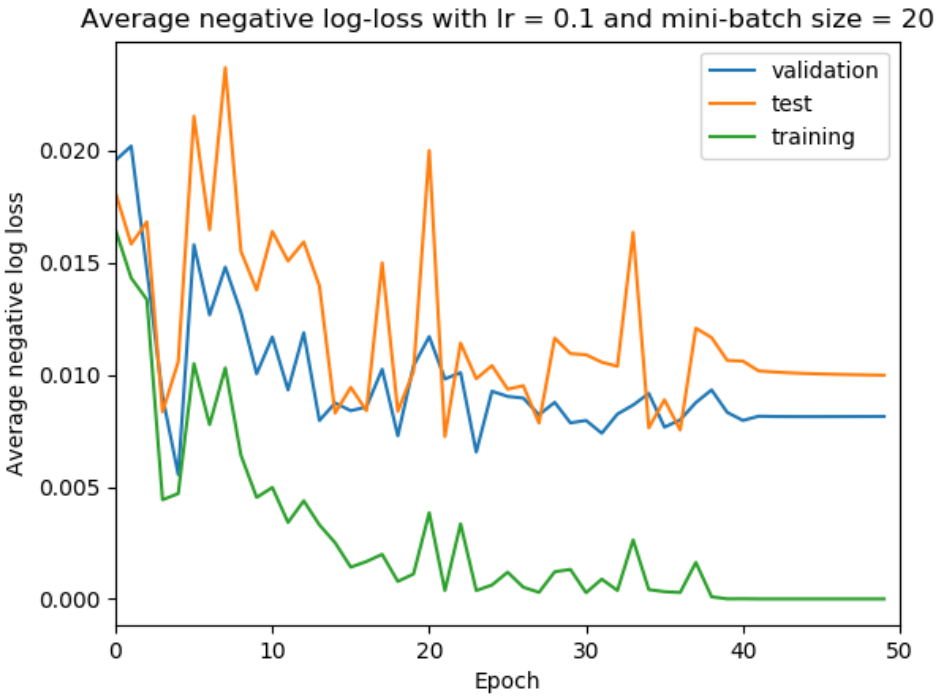
e)

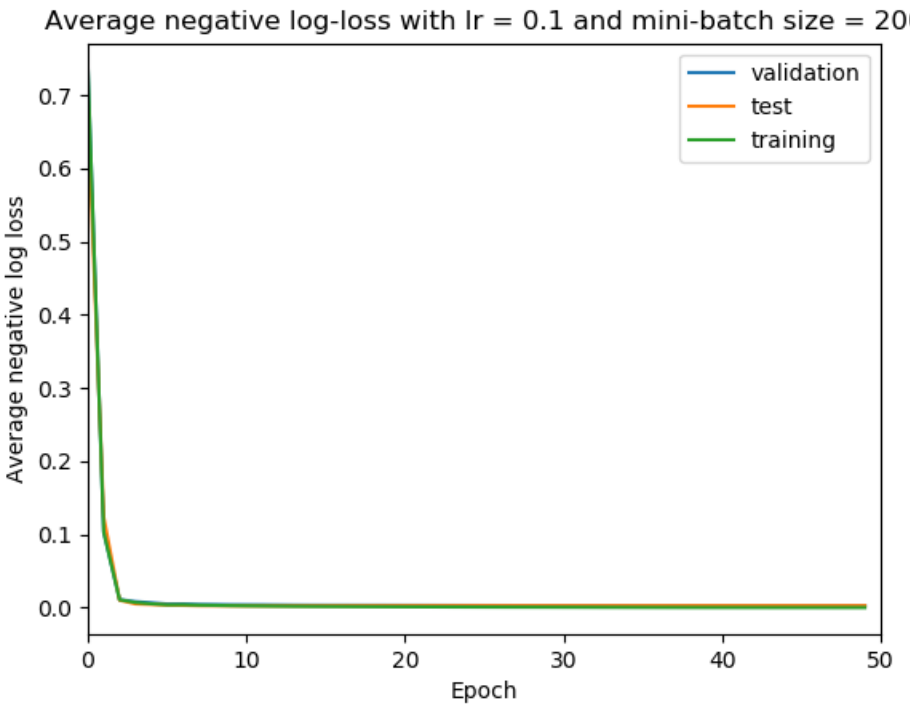
Pr(W=1|P=1) est donnée dans les tables de probabilités conditionnelles, il suffit uniquement d'aller la chercher dans le vecteur Watson.

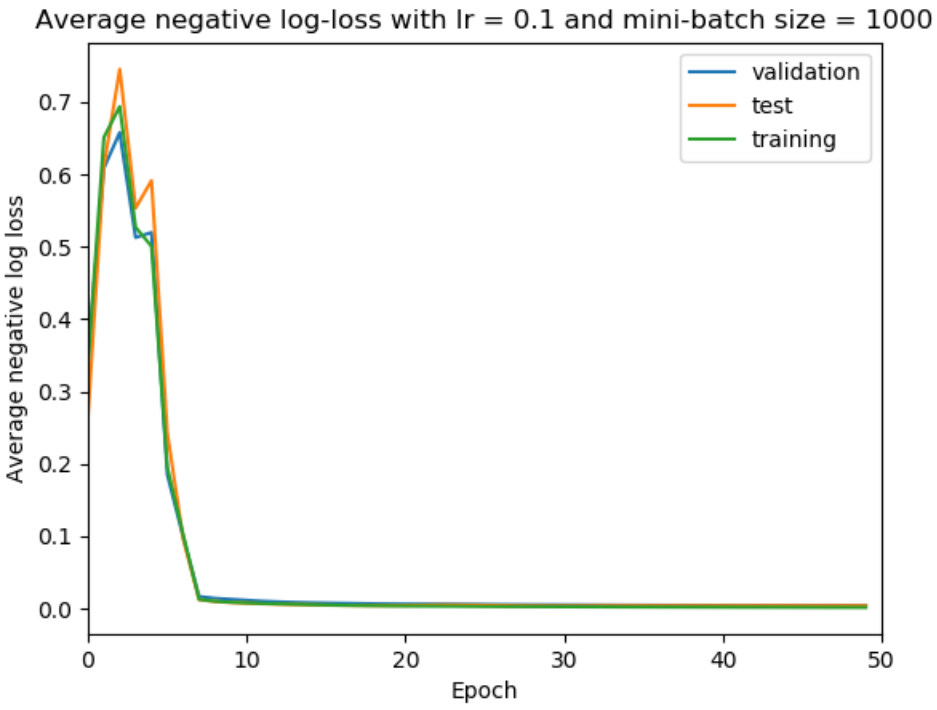
## 2.3

a)

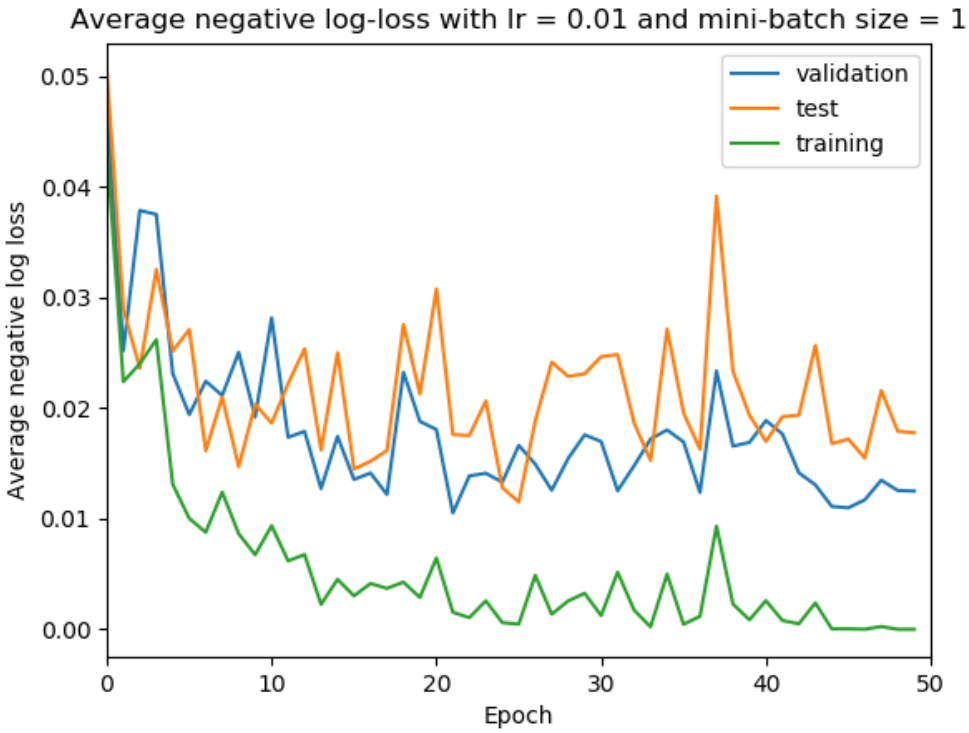
**lr = 0.1**

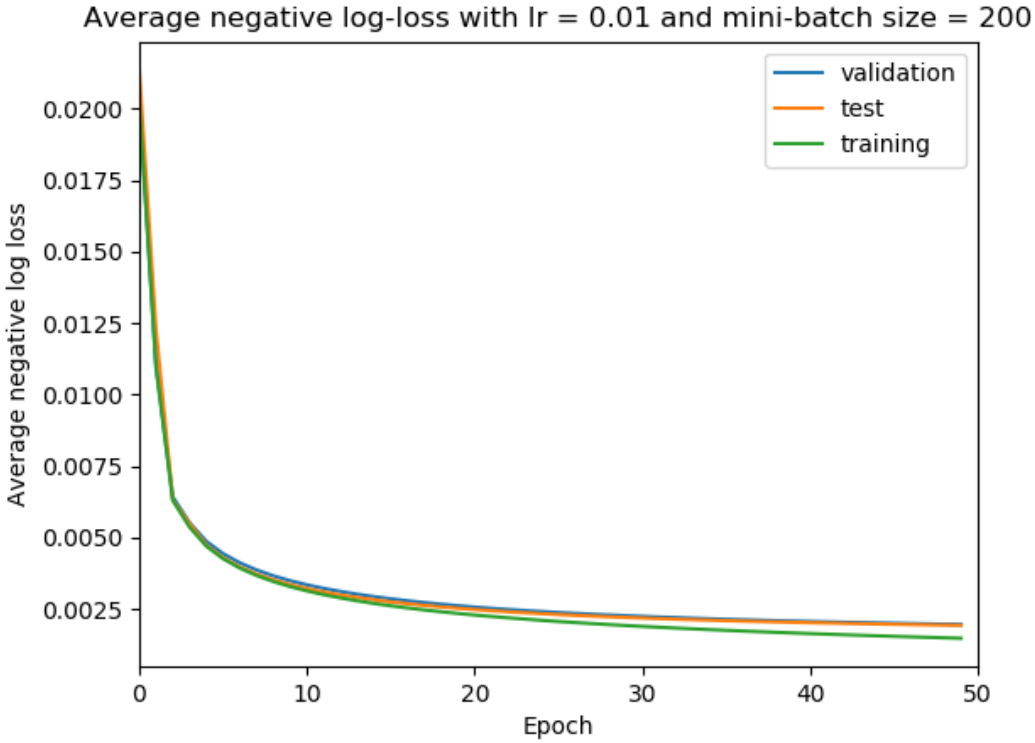


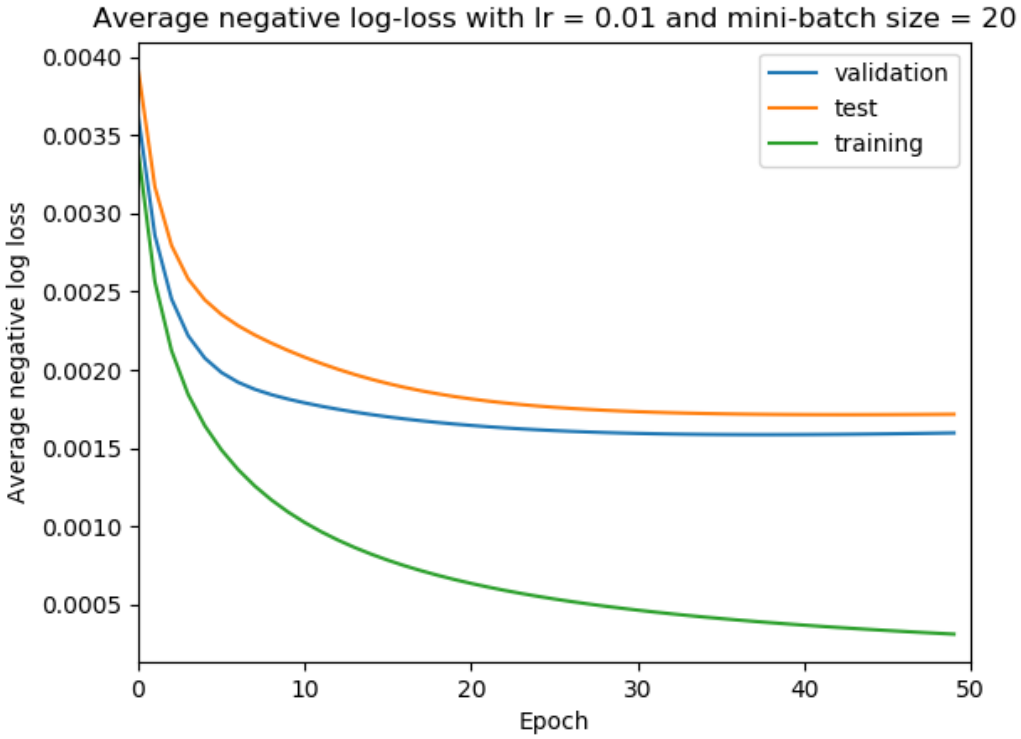

Average negative log-loss with lr = 0.1 and mini-batch size = 200
validation
test
training
Average negative log loss
Epoch
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

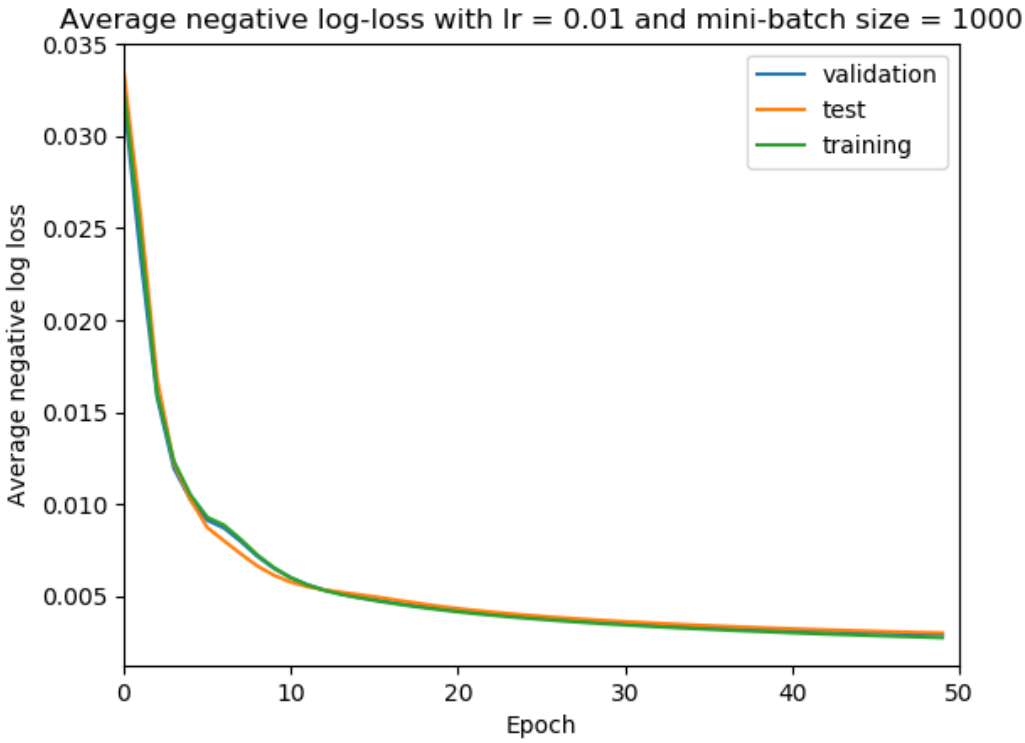

Average negative log-loss with lr = 0.1 and mini-batch size = 1000
validation
test
training
Average negative log loss
Epoch
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

**lr = 0.01**

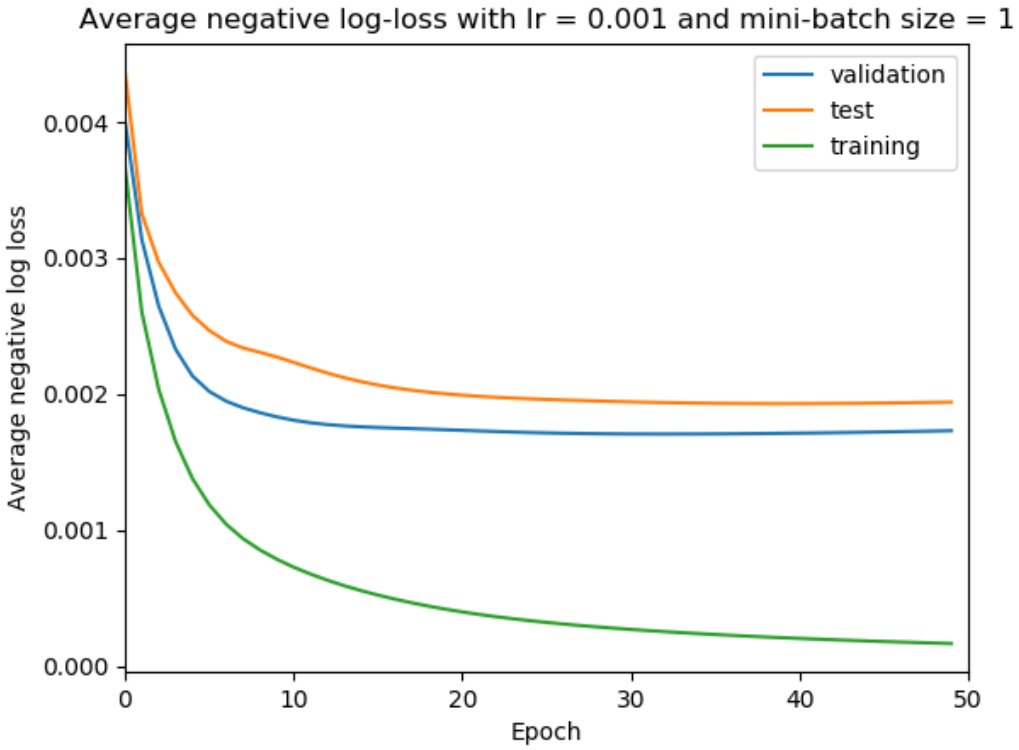

Average negative log-loss with lr = 0.01 and mini-batch size = 1
validation
test
training
Average negative log loss
Epoch
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50

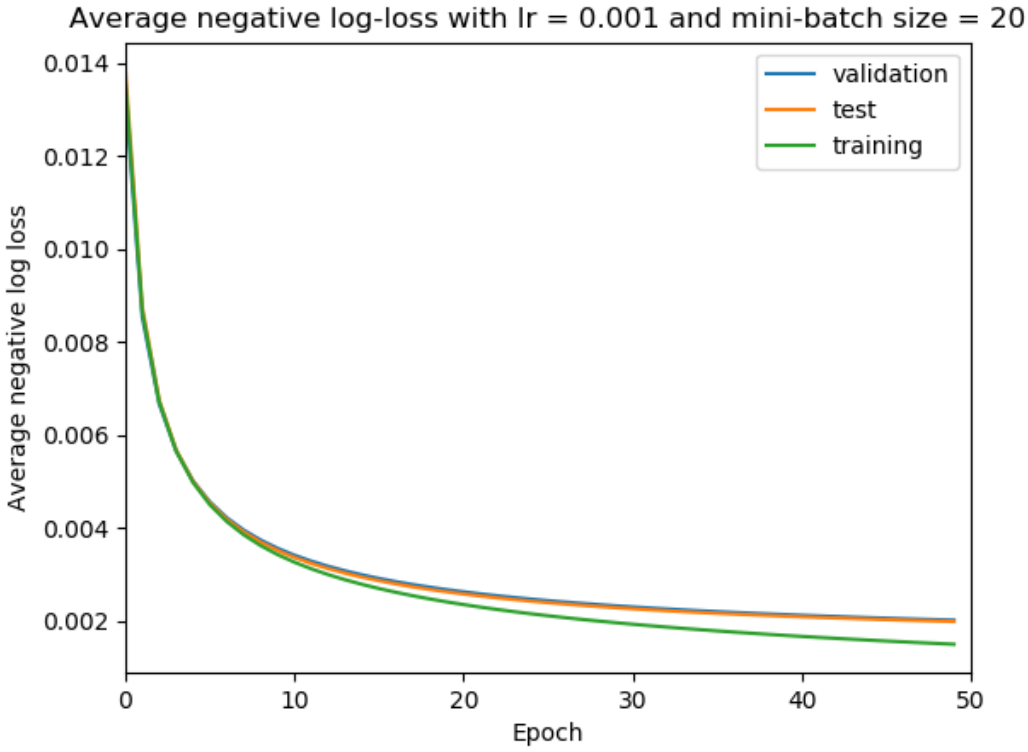

Average negative log-loss with lr = 0.01 and mini-batch size = 200
validation
test
training
Average negative log loss
Epoch
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0
10
20
30
40
50

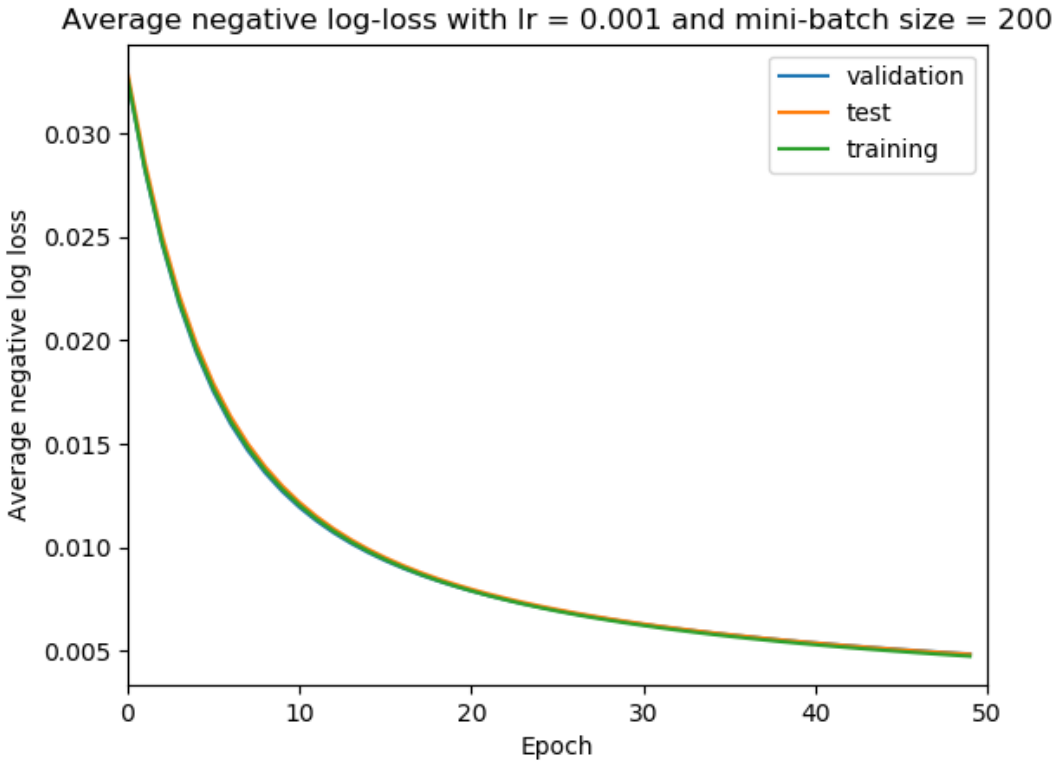

Average negative log-loss with lr = 0.01 and mini-batch size = 20
validation
test
training
Average negative log loss
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0
10
20
30
40
50
Epoch

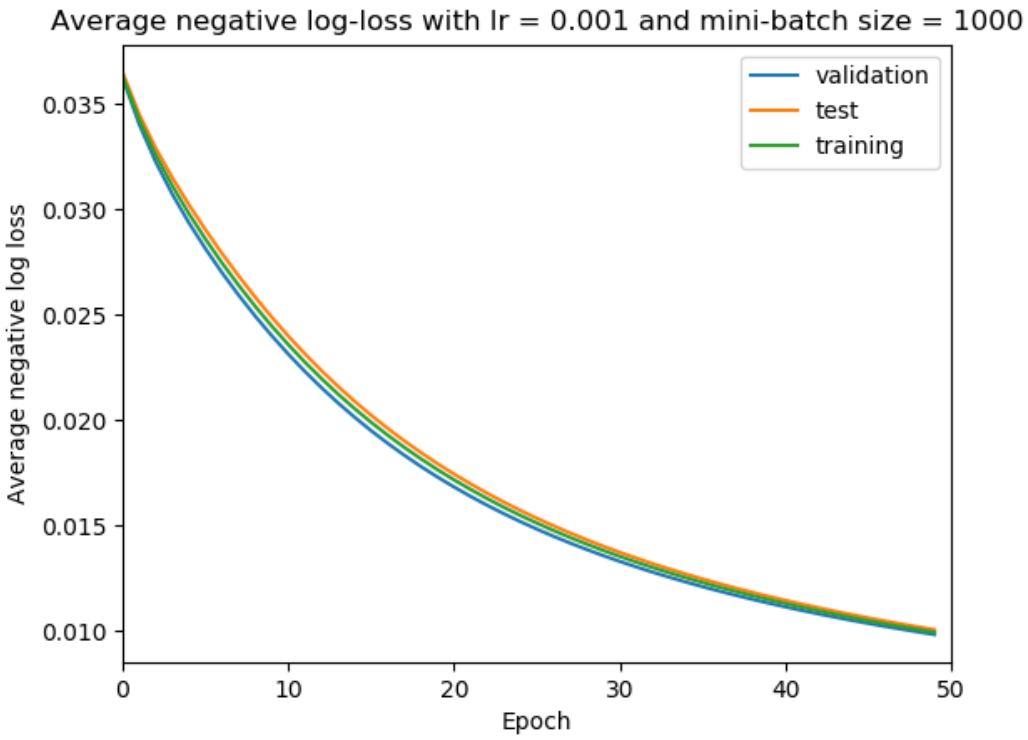

Average negative log-loss with lr = 0.01 and mini-batch size = 1000
validation
test
training
Average negative log loss
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0
10
20
30
40
50
Epoch

**lr = 0.001**


Average negative log-loss with lr = 0.001 and mini-batch size = 1
validation
test
training
Average negative log loss
0.004
0.003
0.002
0.001
0.000
0
10
20
30
40
50
Epoch


Average negative log-loss with lr = 0.001 and mini-batch size = 20
validation
test
training
Average negative log loss
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0
10
20
30
40
50
Epoch


Average negative log-loss with lr = 0.001 and mini-batch size = 200
validation
test
training
Average negative log loss
0.030
0.025
0.020
0.015
0.010
0.005
0
10
20
30
40
50
Epoch


Average negative log-loss with lr = 0.001 and mini-batch size = 1000
validation
test
training
Average negative log loss
0.035
0.030
0.025
0.020
0.015
0.010
0
10
20
30
40
50
Epoch

b)

Nouvelle équation pour le gradient :

$$L(f_i(xi;\ \theta), y_i) = log \prod_{i=1}^{N} P(\tilde{y}_i|\tilde{x}_i;\theta) + \ \lambda_1 * \sum_{w_i} w_i^2 + \ \lambda_2 * \sum_{w_i} |w_i|$$

$$\Rightarrow \frac{\partial L(f_i(xi;\ \theta), y_i)}{\partial \theta} = \sum_{i=1}^{N} \tilde{y}_i \tilde{x}_i{}^T - \sum \hat{y}_i \tilde{x}_i{}^T + \ \lambda_1 * 2W + \ \lambda_2 * \frac{W}{\|W\|}$$

Tableau#1 : variance et moyenne des dimensions aléatoires et non-aléatoires avec régularisation pour lr = 0.001, taille mini-batch = 1, l1 = 0.01, l2= 0.01, nombre de répétition = 3

| | Moyenne | Variance |
|---|---|---|
| Dimensions aléatoires | -9.565495e-05 | 3.974653e-03 |
| Dimensions non-aléatoires | -1.142939e-04 | 1.572114e-02 |
| Rapport | 1.194856 | 3.955349 |

Tableau#2 : variance et moyenne des dimensions non-aléatoires sans régularisation pour lr = 0.001, taille mini-batch = 1, nombre de répétition = 3

| | Moyenne | Variance |
|---|---|---|
| Dimensions non-aléatoires | -3.031088e-04 | 1.674570e-02 |

<u>Comparaison Tableau#1</u>

On peut remarquer dans le tableau#1 que la variance pour les dimensions non-aléatoires avec régularisation est significativement plus élevée que la variance pour les dimensions aléatoires (3.955349 plus élevé). La régularisation explique une partie du phénomène. Effectivement, la régularisation L1 pousse les poids vers 0. Puisque les poids des dimensions aléatoires ne servent pas pour le calcul de la classe, ils sont uniquement déplacés par la régularisation qui diminue leur variance.

Comparaison Tableau#2 par rapport au Tableau#1

La moyenne des poids associés aux dimensions non-aléatoires d'un entrainement non régularisé est presque 3 fois celle des poids pour un entrainement régularisé. L'objectif de la régularisation est de réduire les poids du modèle pour s'assurer que le modèle ne devienne pas trop biaisé (surentrainement). Il n'est donc pas étonnant d'observer ce ratio.

**Pouvez-vous trouver une combinaison de valeurs pour les deux termes de régularisation qui ramènent à zéro les poids associés aux dimensions aléatoires ajoutées?**

Une combinaison de valeurs pour les deux termes de régularisation qui ramènent à zéro les poids associés aux dimensions aléatoires ajoutés peuvent probablement être trouvé. Ces valeurs seraient probablement assez grandes pour permettre à la mise à jour des poids d'écraser complètement les poids des dimensions aléatoires.