Rapport FTML

Hao YE Arthur FAN Pierre-Louis LANDOUZI

June 17, 2022

Contents

1 BAYES ESTIMATOR AND BAYES RISK

1.1 Question 1

On pose $X \sim Binomial(2, \frac{1}{4})$, $Y = \begin{cases} B(1/3) & \text{Si X = 0} \\ B(2/3) & \text{Si X = 1} \\ B(1/4) & \text{Si X = 2} \end{cases}$, $\mathcal{X} = \{0, 1, 2\}$ et $\mathcal{Y} = \{0, 1\}$.
Pour la fonction de loss, on choisit la fonction $0 - 1 loss$
On rappelle la définition du bayes prédicteur $f^*(x)$

$$\begin{aligned} f^*(x) &= \underset{y \in \mathcal{Y}}{argmin}(E(l(Y, y)|X = x)) \\ &= \underset{y \in \mathcal{Y}}{argmin}(P(Y \neq y|X = x)) \\ &= \underset{y \in \mathcal{Y}}{argmin}(1 - P(Y = y|X = x)) \\ &= \underset{y \in \mathcal{Y}}{argmin}(P(Y = y|X = x)) \end{aligned} \tag{1}$$

Ainsi $f^*(x) = \begin{cases} 0 & \text{Si x = 0} \\ 1 & \text{Si x = 1} \\ 0 & \text{Si x = 2} \end{cases}$
On rapelle la définition du risque de bayes R^* avec $\eta(x) = P(Y = 1|X = x)$

$$\begin{aligned} R^* &= E[min(\eta(x), 1 - \eta(x))] \\ &= P(X = 0) * min(\eta(0), 1 - \eta(0)) \\ &+ P(X = 1) * min(\eta(1), 1 - \eta(1)) \\ &+ P(X = 2) * min(\eta(2), 1 - \eta(2)) \end{aligned} \tag{2}$$

Comme $X \sim Binomial(2, 1/4)$, ainsi

$$\begin{aligned} P(X = 0) &= \frac{9}{16} \\ P(X = 1) &= \frac{3}{8} \\ P(X = 2) &= \frac{1}{16} \end{aligned} \tag{3}$$

Et pour $\eta(x)$:

$$\begin{aligned} \eta(0) &= P(Y = 1|X = 0) = \frac{1}{3} \\ \eta(1) &= P(Y = 1|X = 1) = \frac{2}{3} \\ \eta(2) &= P(Y = 1|X = 2) = \frac{1}{4} \end{aligned} \tag{4}$$

Donc pour R^*

$$\begin{aligned} R^* &= \frac{9}{16} * \frac{1}{3} + \frac{3}{8} * \frac{1}{3} + \frac{1}{16} * \frac{1}{4} \\ &= \frac{21}{64} \\ &= 0.328125 \end{aligned} \tag{5}$$

1.2 Question 2

On propose $\tilde{f}(x) = \begin{cases} 1 & \text{Si x = 0} \\ 0 & \text{Si x = 1} \\ 1 & \text{Si x = 2} \end{cases}$.

La formule du risque général est $R(\tilde{f})$

$$\begin{aligned} R(\tilde{f}) &= E(l(Y, \tilde{f}(X))) \\ &= 1 * P(Y \neq \tilde{f}(X)) \\ &= P((Y \neq \tilde{f}(X)) \cap (X = 0)) \\ &+ P((Y \neq \tilde{f}(X)) \cap (X = 1)) \\ &+ P((Y \neq \tilde{f}(X)) \cap (X = 2)) \\ &= P(X = 0) * P((Y \neq \tilde{f}(X)|X = 0) \\ &+ P(X = 1) * P((Y \neq \tilde{f}(X)|X = 1) \\ &+ P(X = 2) * P((Y \neq \tilde{f}(X)|X = 2) \\ &= \frac{9}{16} * (1 - p) + \frac{3}{8} * q + \frac{1}{16} * (1 - r) \\ &= \frac{9}{16} * \frac{2}{3} + \frac{3}{8} * \frac{2}{3} + \frac{1}{16} * \frac{3}{4} \\ &= \frac{43}{64} \\ &= 0.671875 \end{aligned} \tag{6}$$

On peut également utiliser $R(\tilde{f}) = 1 - R(f^*)$ car $\tilde{f}(x) = 1 - f^*(x)$.Ansin $R(\tilde{f}) = 1 - \frac{21}{64} = \frac{43}{64}$. Nous avons retrouvé le même résultat qu'avec la formule.

2 BAYES RISK WITH ABSOLUTE LOSS

2.1 Question 1

On prend $\mathcal{X} \in \mathbb{R}$ et $\mathcal{Y} \in \mathbb{R}$ et pour X = 0, $Y \sim \mathcal{U}(\{-3, -2, -1, 0, 4\})$
Soient $f_1^* = \underset{z \in \mathcal{Y}}{argmin} E[|y - z||X = 0]$ et $f_2^* = \underset{z \in \mathcal{Y}}{argmin} E[(y - z)^2|X = 0]$

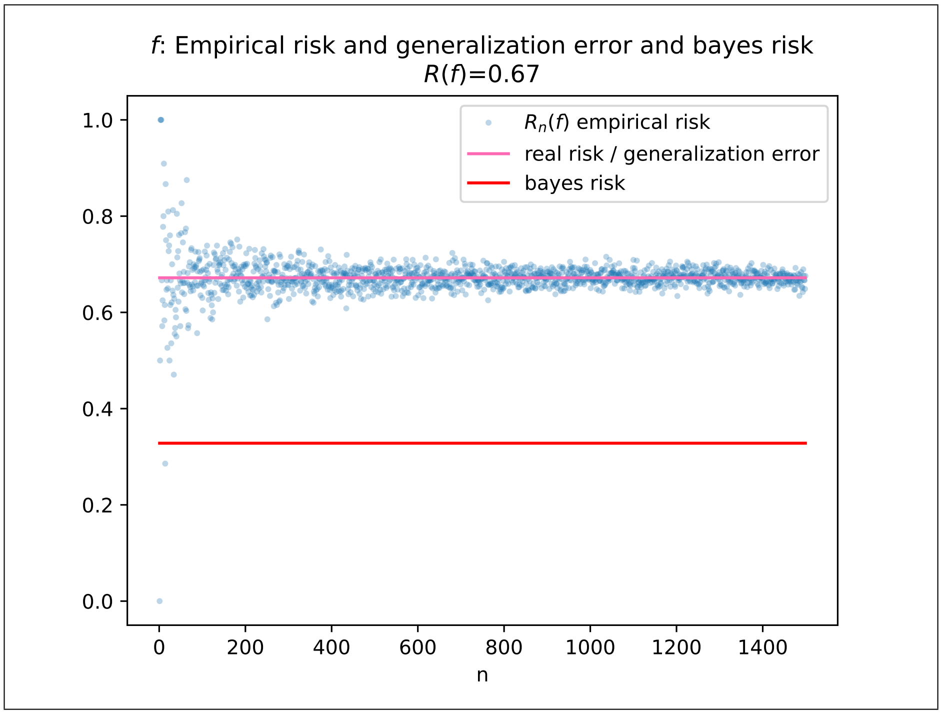

Figure 1: Risque empirique, risque réelle, et risque de bayes

On calcule d'abord pour f_1^*

$$\begin{aligned} z = -3, E[|y-z||X=0] = \frac{1}{5}(|-3+3| + |-2+3| + \\ |-1+3| + |0+3| + |4+3|) = \frac{13}{5} \end{aligned} \tag{7}$$

$$\begin{aligned} z = -2, E[|y-z||X=0] = \frac{10}{5} \\ z = -1, E[|y-z||X=0] = \frac{9}{5} \\ z = 0, E[|y-z||X=0] = \frac{10}{5} \\ z = 4, E[|y-z||X=0] = \frac{22}{5} \end{aligned} \tag{8}$$

Donc $f_1^* = \underset{z \in \mathcal{Y}}{argmin} E[|y-z||X=0] = -1$
On calcule maintenant f_2^*

$$\begin{aligned} z = -3, E[|y-z||X=0] = \frac{1}{5}((-3+3)^2 + (-2+3)^2 + \\ (-1+3)^2 + (0+3)^2 + (4+3)^2) = \frac{63}{5} \end{aligned} \tag{9}$$

$$\begin{aligned}
z=-2, E[|y-z||X=0] &= \frac{42}{5}\\
z=-1, E[|y-z||X=0] &= \frac{31}{5}\\
z=0, E[|y-z||X=0] &= \frac{30}{5}\\
z=4, E[|y-z||X=0] &= \frac{126}{5}
\end{aligned} \tag{10}$$

Donc $f_2^* = \underset{z\in\mathcal{Y}}{argmin} E[|y-z||X=0] = 0$

2.2 Question 2

Soit $f^*(x)$

$$\begin{aligned}
f^*(x) &= \underset{z\in\mathbb{R}}{argmin}(E[|y-z||X=x])\\
&= \underset{z\in\mathbb{R}}{argmin}(g(z))
\end{aligned} \tag{11}$$

Soit $g(x)$

$$g(z) = \int\limits_{y\in\mathbb{R}} (|y-z| \underset{Y|X=x}{p(y)})dy \tag{12}$$

Je pose $\underset{Y|X=x}{p(y)} = f(y|X=x)$

$$\begin{aligned}
g(z) = &\int\limits_{y\in\mathbb{R}} (|y-z|f(y|X=x))dy\\
&= \int_{-\infty}^{z} (z-y)f(y|X=x)dy + \int_{z}^{+\infty} (y-z)f(y|X=x)dy\\
&= zF(z|X=x) - \int_{-\infty}^{z} yf(y|X=x)dy - z(1-F(z|X=x))\\
&+ \int_{z}^{+\infty} yf(y|X=x)dy\\
&= \int_{z}^{+\infty} yf(y|X=x)dy - z(1-F(z|X=x)) + zF(z|X=x)\\
&- \int_{-\infty}^{z} yf(y|X=x)dy
\end{aligned} \tag{13}$$

On dérive $g(z)$ par z

$$\begin{aligned}
\frac{dg}{dz}(z) &= -zf(z|X=x) - 1 + F(z|X=x) + zf(z|X=x)\\
&+ F(z|X=x) + zf(z|X=x) - zf(z|X=x)\\
&= 2 * F(z|X=x) - 1
\end{aligned} \tag{14}$$

Pour trouver le minimum, on pose $\frac{dg}{dz}(z) = 0$

$$\begin{aligned}
\frac{dg}{dz}(z) &= 0 \\
&\Rightarrow 2 * F(z|X = x) = 1 \\
&\Rightarrow F(z|X = x) = \frac{1}{2}
\end{aligned} \tag{15}$$

Le bayes prédicteur $f^*(x)$ est la médiane.

3 EXPECTED VALUE OF EMPIRICAL RISK

3.1 Question 1

$$\begin{aligned}
E[R_n\theta] &= E[\frac{1}{n}||y - X\theta||_2^2]) \\
&= E[\frac{1}{n}||X\theta^* + \varepsilon - X\theta||_2^2]) \\
&= E[\frac{1}{n}||X\theta^* + \varepsilon - X(X^TX)^{-1}X^Ty||_2^2]) \\
&= E[\frac{1}{n}||X\theta^* + \varepsilon - X(X^TX)^{-1}X^T(X\theta^*+)||_2^2]) \\
&= E[\frac{1}{n}||X\theta^* + \varepsilon - X(X^TX)^{-1}X^TX\theta^* + X(X^TX)^{-1}X^T||_2^2]) \\
&= E[\frac{1}{n}||X\theta^* + \varepsilon - XI_n\theta^* + X(X^TX)^{-1}X^T\varepsilon||_2^2]) \\
&= E[\frac{1}{n}||\varepsilon - X(X^TX)^{-1}X^T\varepsilon||_2^2]) \\
&= E[\frac{1}{n}||(I_n - X(X^TX)^{-1}X^T)\varepsilon||_2^2])
\end{aligned} \tag{16}$$

3.2 Question 2

$$\begin{aligned}
(A^TA) = \sum_{k=1}^{n} A_{ik}A_{kj}^T &= \sum_{k=1}^{n} A_{ik}A_{jk} \\
tr(A^TA) &= \sum_{(ij)\in[1,n]^2} A_{ij}^2
\end{aligned} \tag{17}$$

3.3 Question 3

$$\begin{aligned} E[\frac{1}{n}||(I_n - X(X^TX)^{-1}X^T)\varepsilon||^2] &= E[\frac{1}{n}||A\varepsilon||^2] \\ &= E[\frac{1}{n}\sum_{i=1}^{n}(A_i\varepsilon)^2] \\ &= E[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon^TA_i^T)^2] \\ &= E[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon^TA_i^TA_i\varepsilon)] \\ &= E[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon^T\varepsilon A_i^TA_i)] \\ &= E[\frac{1}{n}\varepsilon^T\varepsilon\sum_{i=1}^{n}(A_i^TA_i)] \\ &= E[\frac{1}{n}\varepsilon^2\sum_{(ij)\in[1,n]^2}(A_{ij}^2)] \\ &= \frac{1}{n}\sigma^2Tr(A^TA) \end{aligned} \tag{18}$$

3.4 Question 4

$$A = In - X(X^TX)^{-1}X^T \tag{19}$$

$$\begin{aligned} A^T &= (In - X(X^TX)^{-1}X^T)^T \\ &= In^T - X((X^TX)^{-1})^TX^T \\ &= In^T - X(X^TX)^{-1}X^T \end{aligned} \tag{20}$$

$$\begin{aligned} A^TA &= (In^T - X(X^TX)^{-1}X^T)(In - X(X^TX)^{-1}X^T) \\ &= In - 2(X(X^TX)^{-1}X^T) + (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\ &= In - 2X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= In - 2X(X^TX)^{-1}X^T + XIn(X^TX)^{-1}X^T \\ &= In - X(X^TX)^{-1}X^T \\ &= A \end{aligned} \tag{21}$$

3.5 Question 5

Nous savons que la matrice X est de taille (n,d) Grâce aux questions précédentes nous pouvons trouver:

$$\begin{aligned} E[R_n\hat{\theta}] &= E[\frac{1}{n}||(In - X(X^TX)^{-1}X^T)\varepsilon||_2^2]) \\ &= E[\frac{1}{n}||A\varepsilon||_2^2]) \\ &= \frac{1}{n}\sigma^2 Tr(A^TA) \\ &= \frac{1}{n}\sigma^2 Tr(A) \end{aligned} \tag{22}$$

$$\begin{aligned} A &= In - X(X^TX)^{-1}X^T \\ Tr(A) &= Tr(In - X(X^TX)^{-1}X^T) \\ &= Tr(In) - Tr(X(X^TX)^{-1}X^T) \\ &= n - Tr(X(X^TX)^{-1}X^T) \\ &= n - Tr(X^TX(X^TX)^{-1}) \\ &= n - Tr(Id) \\ &= n - d \end{aligned} \tag{23}$$

Donc:

$$\begin{aligned} E[Rn\hat{\theta}] &= \frac{1}{n}\sigma^2 Tr(A) \\ &= \frac{1}{n}\sigma^2(n-d) \\ &= \frac{n-d}{n}\sigma^2 \end{aligned} \tag{24}$$

3.6 Question 6

On veut montrer que

$$E[Rx(\hat{\theta})] = \frac{n-d}{n}\sigma^2 \tag{25}$$

On pose:

$$\begin{aligned} gn(\theta) &= \frac{||y - X\hat{\theta}||_2^2}{n-d} \\ &= \frac{nRn(\hat{\theta})}{n-d} \end{aligned} \tag{26}$$

$$\begin{aligned} E[gn(\hat{\theta})] &= E[\frac{nRn(\hat{\theta})}{n-d}] \\ &= \frac{n}{n-d}\frac{n-d}{n}\sigma^2 \\ &= \sigma^2 \end{aligned} \tag{27}$$

3.7 Question 7

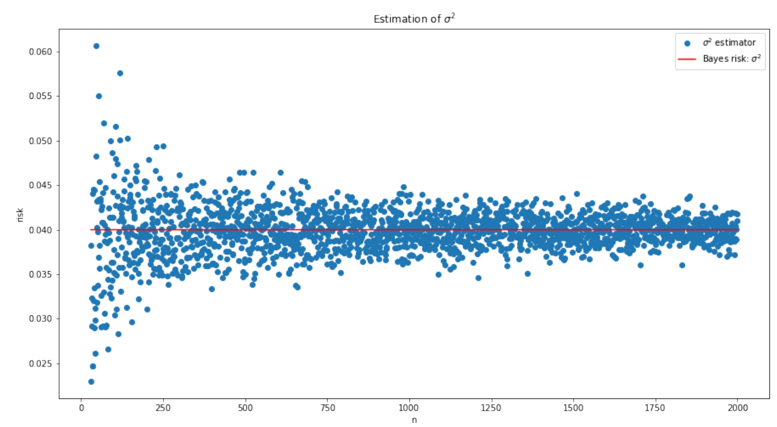

Figure 2: Estimation de σ^2

Nous avons généré plusieurs matrices de X et de Y auxquelles nous avons calculé $\hat{\theta}$. Nous avons fixé la taille de nos matrices et nous avons fait varier les n de 30 à 2000. Puis nous avons estimé les σ^2. On peut ainsi remarquer que plus n augmente, plus notre estimation de σ^2 tend vers la valeur réelle de σ^2.

4 REGRESSION

Pour cette partie, nous avons utilisé trois méthodes de régression:

- Ridge regression
- Lasso regression
- OLS

4.1 Ridge Regression

Pour réaliser le modèle de Ridge, nous avons utilisé le Ridge() du sklearn. Pour trouver le meilleur hyper-paramètre pour notre modèle Ridge regression, nous avons utilisé la fonction GridSearchCV avec la technique de Cross Validation. Nous avons varié α entre 10^{-8} et 10^{5} avec un pas de 0.2. Enfin, nous avons trouvé pour $\alpha = 10$, notre modèle a atteint sa meilleure performance sur le jeu de données d'entraînement, et il obtient un score R^2 de 0.9135 sur le jeu de données de test.

4.2 Lasso Regression

Pour réaliser le modèle Lasso, nous avons procédé de la même manière que Ridge. Nous avons utilisé Lasso() et GridSearchCV() du sklearn. Nous avons varié α entre 10^{-8} et 10^{5} avec un pas de $10^{0.2}$. Pour notre modèle Lasso, son meilleur hyper-paramètre est $\alpha = 1$, et il obtient un score R^2 de 0.9137 sur le jeu de données de test.

4.3 OLS

Pour OLS, nous avons repris la formule que nous avons vue en cours de FTML. L'estimateur OLS $\hat{\theta} = (X^T X)^{-1} X^T Y$, et $Y_{pred} = X\hat{\theta}$. Avec l'estimateur OLS, nous avons obtenu un score R^2 de 0.9124 sur le jeu de données de test.

Modèle	hyperparamètre	score R^2
Ridge regression	$\alpha = 10$	0.9135
Lasso regression	$\alpha = 1$	0.9137
OLS	None	0.9124

5 CLASSIFICATION

Pour cette partie, nous avons utilisé trois méthodes de classification:

- KNN
- Régression logistique
- SVC

5.1 KNN

Pour réaliser le KNN, nous avons repris l'algorithme que nous avons vu en cours de PTML. Pour déterminer le meilleur nombre de voisins pour notre modèle sur ce jeu de données. Nous avons testé un certain nombre de hyper-paramètres différents, et nous avons calculé la moyenne des erreurs au carré pour chaque hyper-paramètre distincte. Pour se faire, nous avons varié k entre 0 et 100. Et nous avons trouvé qu'avec $k = 23$, notre modèle a minimisé l'erreur, et

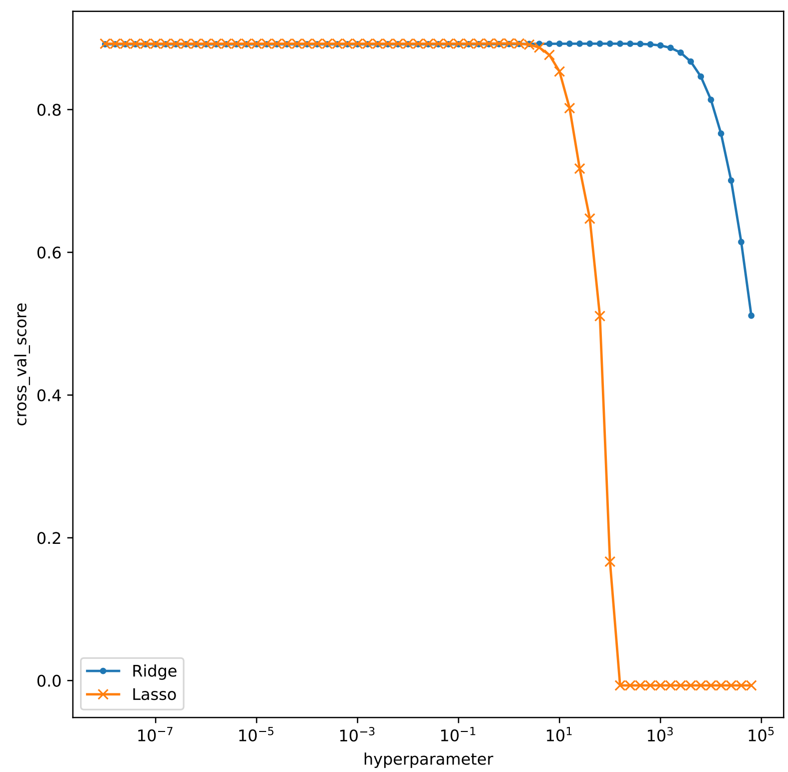

Figure 3: L'évolution du score en fonction du hyperparamètre α pour Ridge et Lasso

il obtient une accuracy de 0.86 sur le jeu de données de test. Cependant, nous avons également utilisé le KNN et GridSearchCV() du sklearn. Avec le KNN et le GridSearchCV() du sklearn, nous avons obtenu une accurracy de 0.836 avec $k = 53$.

5.2 Régression logistique

Pour réaliser la régression logistique, nous avons utilisé la LogisticRegression() du sklearn. Et nous avons garder les paramètre par défaut du sklearn. À l'aide de la LogisticRegression(), nous avons obtenu une accurracy de 0.896 sur le jeu de données de test.

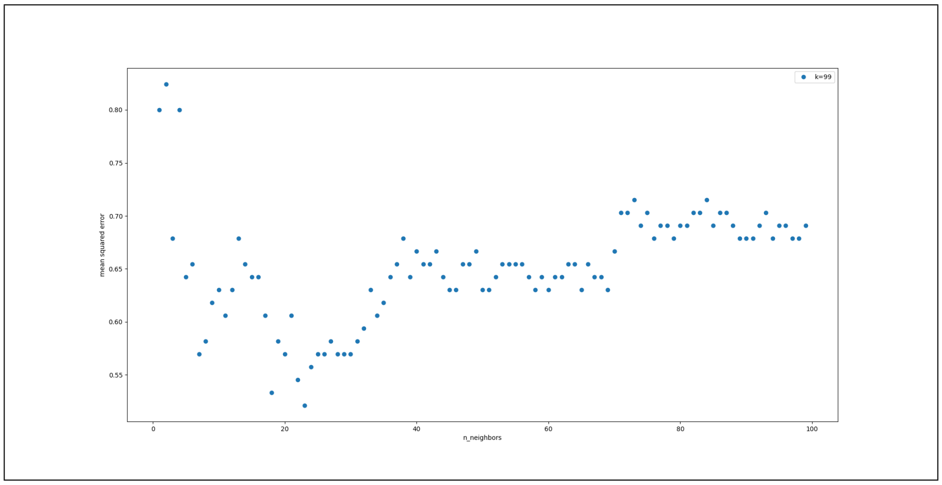

Figure 4: L'erreur en fonction du nombre de voisin

5.3 SVC

Pour réaliser la SVC, nous avons utilisé la SVC(), et pour trouver les meilleurs hyper-paramètres, nous avons utilisé le GridSearchCV() avec la grille de paramètre suivante :

- C = [0.1,1, 10, 100]
- γ = [1,0.1,0.01,0.001]
- kernel = [rbf, poly, sigmoid, linear]

Notre modèle a obtenu une accurracy de 0.881 le jeu de données de test avec C = 1, gamma = 1, et kernel = linear.

Modèle	hyperparamètre	accuracy
KNN user defined	$k = 23$	0.86
KNN with sklearn	$k = 53$	0.836
Linear Regression	None	0.896
SVC	$C = 1$, $\gamma = 1$,kernel = linear	0.881