Régression Linéaire

Quentin Fournier <quentin.fournier@polymtl.ca>

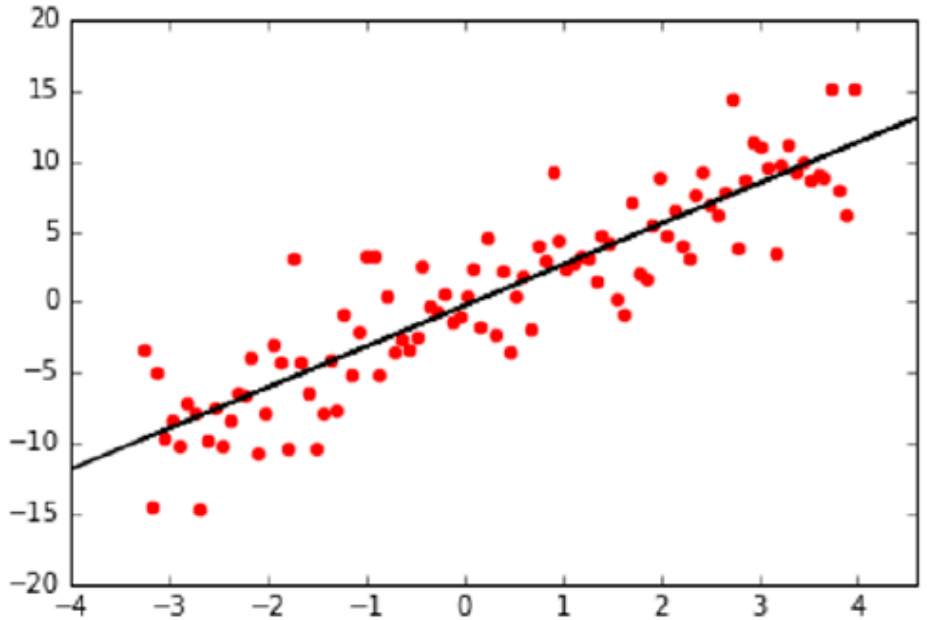
Les diapositives ont été créées par Daniel Aloise
<daniel.aloise@polymtl.ca>

4 octobre 2021

Régression linéaire

- Soient X un ensemble de données et y les valeurs numériques à prédire. La régression linéaire permet de trouver l'hyperplan qui prédit au mieux y_i en fonction de X_i.
- Ex. à une dimension :

y: la valeur qu'on essaye de prédire



Pourquoi des fonctions linéaires ?

- Les relations linéaires sont faciles à comprendre et souvent appropriées en tant que modèle par défaut, ex. :
  - Les prix des logements augmentent linéairement avec la superficie.
  - Le poids augmente linéairement avec les crèmes glacées consommées.

- Si vous voulez qu'une fonction soit linéaire, mesurez-la en deux points seulement ☺

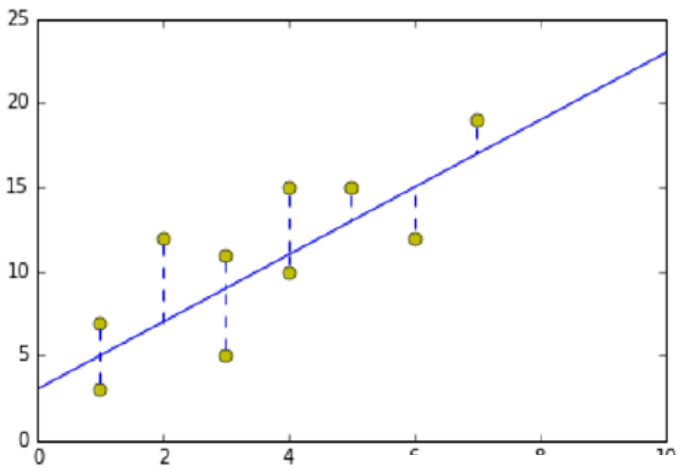
Comment mesure-t-on l'erreur ?

- L'erreur résiduelle est la différence entre les valeurs prédites et les valeurs réelles :

$$r_i = f(X_i) - y_i$$

- Dans le cas de la régression linéaire, on veut minimiser :

$$O = \sum_{i=1}^{n} (f(X_i) - y_i)^2, \quad \text{où} \quad f(X_i) = w_0 + \sum_{j=1}^{d} w^j X_i^j$$



Pourquoi les erreurs au carré ?

On va pénaliser les erreurs qui sont grandes, on veut éviter d'avoir de grandes erreurs

Régression linéaire

- Le coefficient w_0 correspond à l'ordonnée à l'origine (*y-intercept*) de la droite de la régression.
- On n'a pas besoin de w_0 si les données sont centrées (exercice sur Moodle).

Régression linéaire

- On peut réécrire :

$$O = \sum_{i=1}^{n} (w^\top X_i - y_i)^2 = \|Xw - y\|^2$$

- Le gradient de O par rapport à W est égal au vecteur $2X^T(Xw - y)$
- En mettant le gradient égal à zéro, on obtient les conditions d'optimisation suivantes :

$$X^T X w = X^T y$$

- Dans le cas où $X^T X$ est inversible, $w = (X^T X)^{-1} X^T y$

EXAM

Régression linéaire

À noter

$\|x\|^2 = x^T x = \sum_x x_i^2$

$Xw = y$

$$O = \sum_{i=1}^{n} (w^T X_i - y_i)^2 \quad (1)$$

$$= \|Xw - y\|^2 \quad (2)$$

$$= (Xw - y)^T (Xw - y) \quad (3)$$

$$= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw \quad (4)$$

$$= y^T y - 2w^T X^T y + w^T X^T Xw \quad (5)$$

À mettre feuille notes

$$\frac{\partial O}{\partial w} = \frac{\partial}{\partial w}(y^T y - 2w^T X^T y + w^T X^T Xw) \quad (6)$$

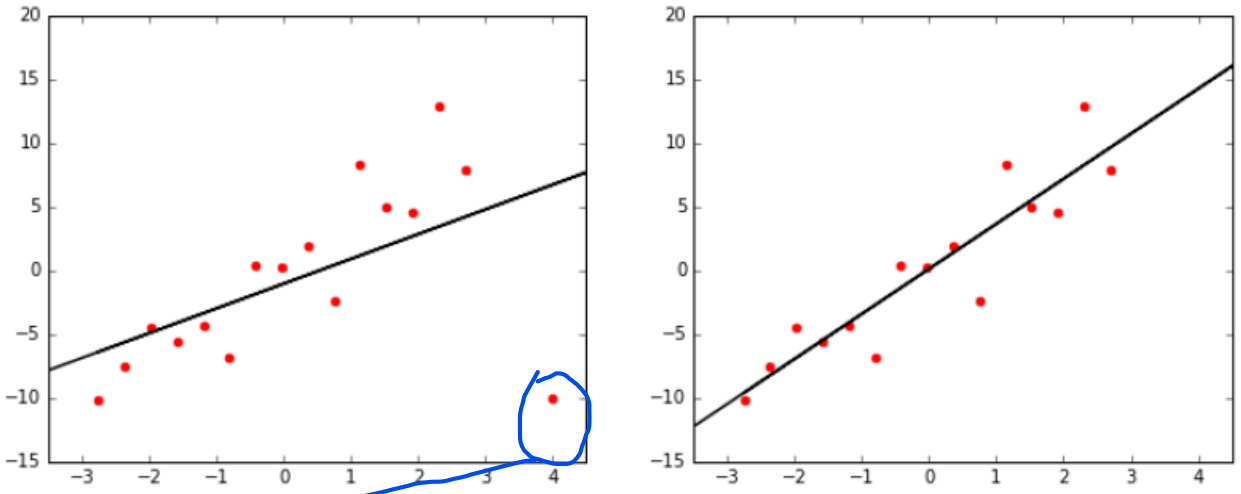
$$= -2X^T y + 2(X^T X)w \quad (7)$$

Régression linéaire - améliorations

- Le traitement approprié des données donne souvent de meilleurs modèles :
  - suppression des valeurs aberrantes (*outliers*)
  - ajustement par des fonctions non linéaires
  - mise à l'échelle des variables dépendantes et indépendantes
  - nettoyage des attributs fortement corrélés

Suppression de valeurs aberrantes

- En raison du poids quadratique des résidus, les points périphériques peuvent affecter largement la régression.
- Identifier les points périphériques et les supprimer de manière raisonnée peut rendre l'ajustement plus robuste.



Skiena, 2017

Régression linéaire est sensible aux données aberrantes. On peut repérer ces valeurs erronnées là et les supprimer

Ajustement des fonctions non linéaires

- La régression linéaire ajuste les données par des droites uniquement !

Ajustement des fonctions non linéaires

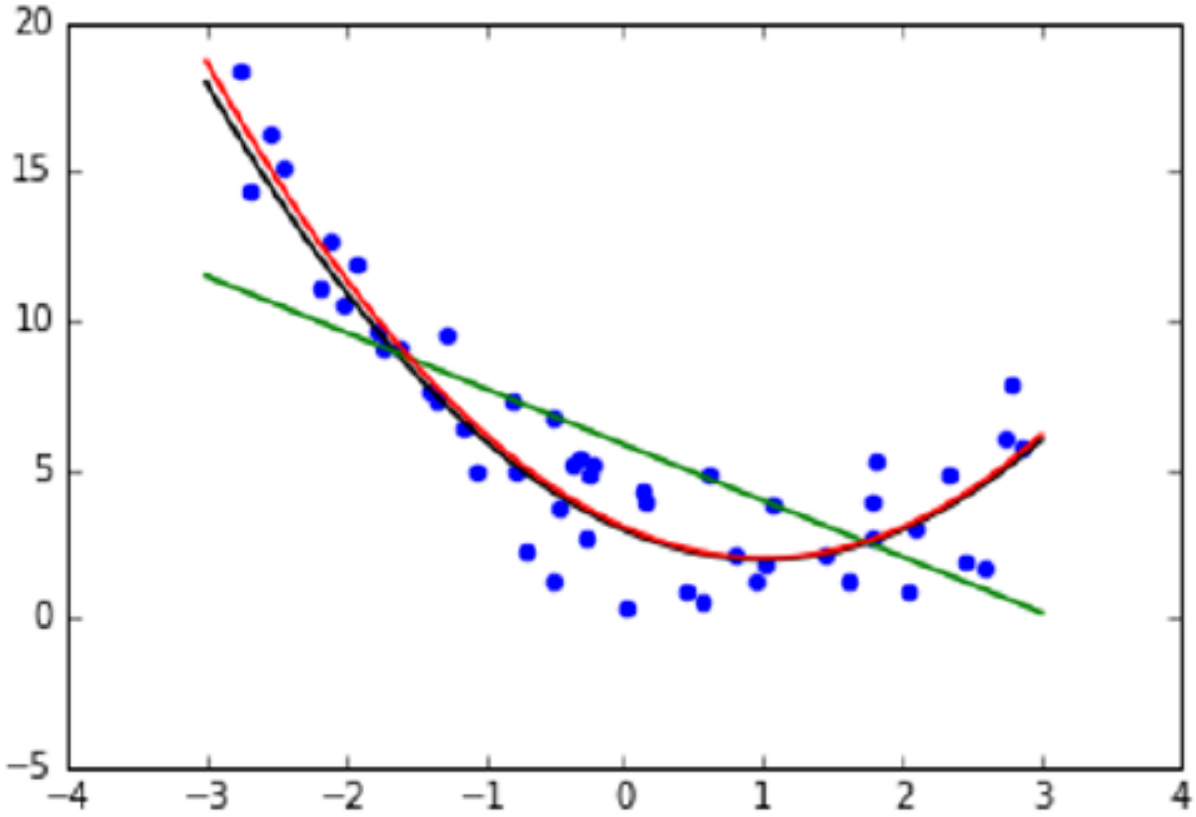
- La régression linéaire ajuste les données par des droites uniquement !
- Pas de soucis : nous pouvons les ajuster aussi par des fonctions quadratiques en créant une autre variable avec la valeur x^2 dans notre matrice de données.

Ajustement des fonctions non linéaires

- La régression linéaire ajuste les données par des droites uniquement !
- Pas de soucis : nous pouvons les ajuster aussi par des fonctions quadratiques en créant une autre variable avec la valeur x^2 dans notre matrice de données.
- Ainsi, le modèle 1-D $f(x) = w_1 x + w_2 x^2$ est quadratique, mais linéaire en w.

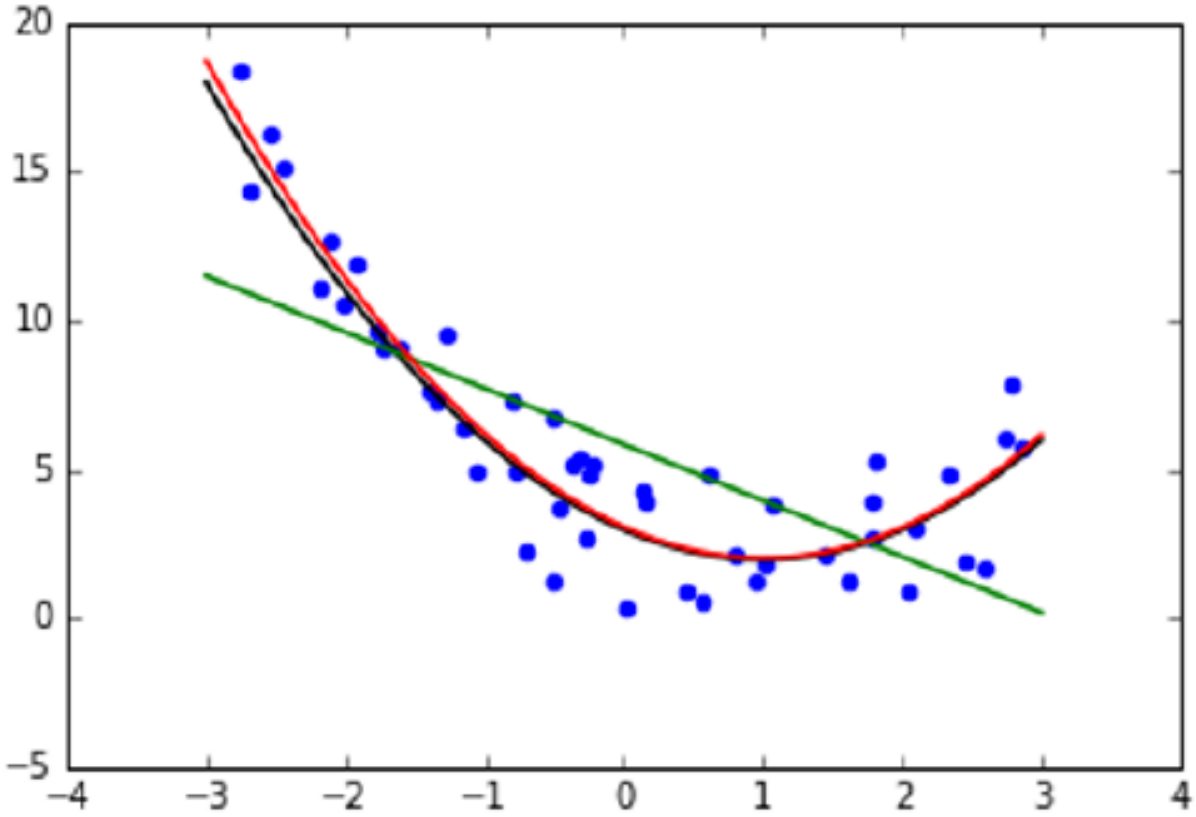
Ajustement des fonctions non linéaires

- La régression linéaire ajuste les données par des droites uniquement !
- Pas de soucis : nous pouvons les ajuster aussi par des fonctions quadratiques en créant une autre variable avec la valeur x^2 dans notre matrice de données.
- Ainsi, le modèle 1-D $f(x) = w_1 x + w_2 x^2$ est quadratique, mais linéaire en w.
- Nous pouvons faire de même avec des fonctions arbitraires en ajoutant explicitement les colonnes correspondantes dans X : ex. $\sqrt{x}, \log x, x^3, 1/x$.



Noire courbe de génération de données
w1x1 nous donne le y

Skiena, 2017



Skiena, 2017

Cependant, l'inclusion explicite de tous les termes non linéaires possibles n'est pas pratique.

Mise à l'échelle

- Des attributs sur de larges intervalles numériques (ex., la population nationale par opposition aux fractions des gens) requièrent des coefficients à différentes échelles pour être rassemblés.
- Ex. $f(x) = 0.03 \times x_1 + 300000000 \times x_2$
- Difficile à savoir quel attribut est le plus important pour la régression.

  ex: x1: 100 000 et x2: 0.1
  Donc on ne sait pas lequel est + important
- Apporte des problèmes de précision.

Mise à l'échelle

- Des attributs sur de larges intervalles numériques (ex., la population nationale par opposition aux fractions des gens) requièrent des coefficients à différentes échelles pour être rassemblés.
- Ex. $f(x) = 0.03 \times x_1 + 300000000 \times x_2$
- Difficile à savoir quel attribut est le plus important pour la régression.
- Apporte des problèmes de précision.
- Solution : normalisez les attributs de votre matrice de données !

Mise à l'échelle sous-linéaire

- Considérons un modèle linéaire qui prédit les années d'éducation d'un enfant basé sur le revenu familiale.

Plus le revenu est élevée, plus l'enfant sera éduqué

Mise à l'échelle sous-linéaire

- Considérons un modèle linéaire qui prédit les années d'éducation d'un enfant basé sur le revenu familiale.
- Selon ce modèle, combien d'années d'études seront prédites pour les enfants de Bill Gates ?

Mise à l'échelle sous-linéaire

- Considérons un modèle linéaire qui prédit les années d'éducation d'un enfant basé sur le revenu familiale.
- Selon ce modèle, combien d'années d'études seront prédites pour les enfants de Bill Gates ?
- La normalisation d'un attribut distribué selon la loi de puissance est inutile parce qu'elle ne fait qu'une transformation linéaire des valeurs de l'attribut.

Mise à l'échelle sous-linéaire

- Un écart énorme entre les valeurs les plus grandes / les plus petites et les médianes signifie qu'aucun coefficient ne peut utiliser la fonction de régression sans exploser sur les grandes valeurs.
- La solution est de remplacer ces attributs x par des attributs sous-linéaires comme $\log x$ et $\sqrt{x}$.
- La normalisation de ces nouveaux attributs est beaucoup plus significative.

Mise à l'échelle sous-linéaire

- Le même est vrai dans le sens opposé.
- Essayer de prédire les grands revenus avec des attributs normalisés est problématique : comment pouvez-vous obtenir $ 100k à partir d'attributs de petite échelle ?
- Considérer le logarithme de grande valeur cible donne de meilleurs modèles.

Mise à l'échelle sous-linéaire

- Autre avantage de l'utilisation des logarithmes :
- On ne peut pas ajuster $f(x) = 20000x_1x_2$ cette fonction par une régression linéaire sur x_1 et x_2.
- Pourtant nous savons que :

$$\log f(x) = \log(20000x_1x_2) = \log(20000) + \log(x_1) + \log(x_2)$$

ainsi, on peut ajuster cette fonction avec $\log x_1, \log x_2$ comme colonnes de la matrice X.

Nettoyage des attributs fortement corrélés

- Supposons que vous avez deux attributs parfaitement corrélés (par exemple, la hauteur en pieds, la hauteur en mètres).
- Les attributs parfaitement corrélés ne fournissent aucune information supplémentaire pour la modélisation (sinon, on les ajouterait successivement pour obtenir un modèle parfait !).
- Mais cela n'est pas l'unique problème...
- On peut obtenir des modèles équivalents se basant sur le premier attribut, sur le deuxième ou sur n'importe quelle combinaison linéaire de deux. Lequel doit-on choisir ?
- Dans ce cas, la matrice $X^T X$ n'est pas inversible ($X^T X$ a des lignes dépendantes).

Nettoyage des attributs fortement corrélés

- Solution : identifiez les attributs hautement corrélés (rappelez-vous la séance passée).
- N'importe quel attribut peut être retiré sans grandes conséquences.
- On peut aussi réduire la dimension de nos données automatiquement par SVD ou PCA.

Régression comme un problème d'optimisation

- Calculer les coefficients par l'équation $w = (X^T X)^{-1} X^T y$ pose encore des problèmes :
  - Calculer l'inverse d'une matrice est une opération coûteuse et qui mène souvent à l'instabilité numérique.
  - La matrice $X^T X$ peut être même singulière.

Régression comme un problème d'optimisation

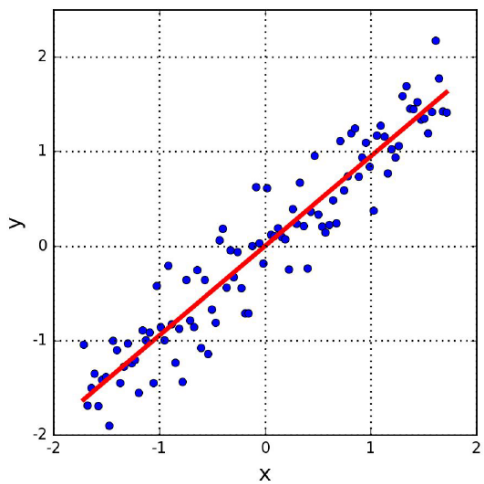
- Nous cherchons des coefficients qui minimisent l'erreur quadratique moyenne des points :

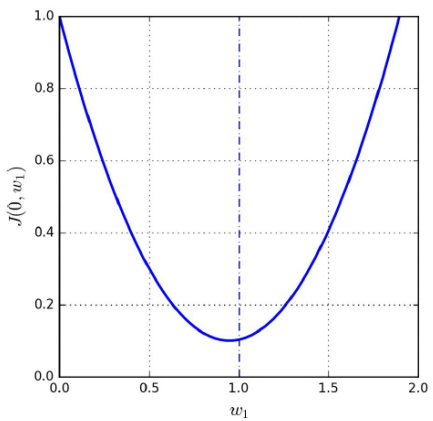
$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - y_i)^2$$

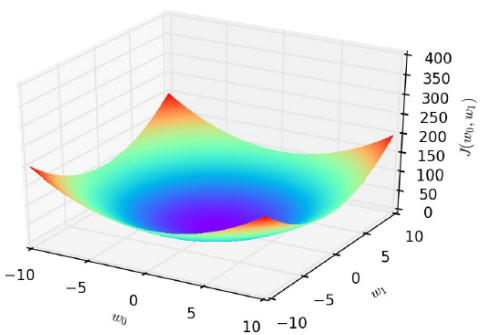
où la droite de la régression est donnée par $f(X_i) = \sum_{j=1}^{d} w^j X_i^j$

Régression comme un problème d'optimisation

- La fonction d'erreur J est convexe :







Skiena, 2017

- Le seul optimum local est un optimum global.

Régression comme un problème d'optimisation

- Quand un espace de recherche est convexe, il est facile de trouver son minimum : continuez simplement à descendre.
- La direction la plus rapide vers le minimum est définie par le gradient.
- Cela motive l'approche de la descente du gradient pour résoudre la régression.

Descente du gradient

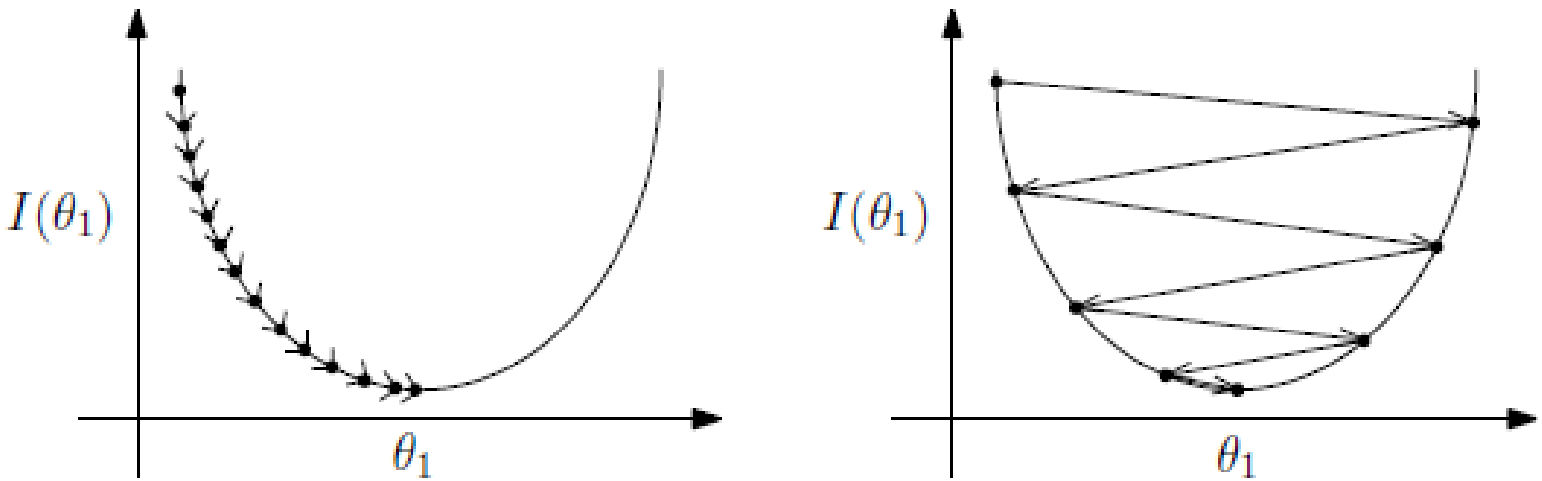
- Le gradient est obtenu par le calcul de la dérivée partielle de $J(w)$ pour chaque coefficient w^j

$$\frac{\partial J}{\partial w^j} = 2 \sum_{i=1}^{n} X_i^j (f(X_i) - y_i)$$

- L'évaluation de chaque dérivée partielle prend un temps linéaire en termes du nombre n d'enregistrements !

Descente du gradient stochastique

- Une bonne heuristique consiste à n'utiliser que quelques enregistrements $n' < n$ pour estimer la dérivée.
- L'ajuste du taux d'apprentissage et de la taille du batch (n') (hyperparamètres) conduit à une optimisation très rapide des fonctions convexes.



Skiena, 2017

Régularisation

- Idéalement, nous aimerions que la régression sélectionne les attributs les plus importants.
- Pourtant, notre fonction objectif ne cherche qu'à minimiser la somme des carrés des erreurs.
- En général, un modèle simple est préférable (c.-à-d. exprimé avec peu d'attributs – rasoir d'Occam).

Régularisation

- Consiste à ajouter des pénalités à la fonction objectif cherchant à garder les coefficients petits

$$J(w) = \sum_{i=1}^{n} (f(X_i) - y_i)^2 + \lambda \sum_{j=1}^{d} (w^j)^2$$

- Cela récompense la mise à zéro des coefficients

Interprétation/Pénalisation

- La nomenclature de la régularisation dépend de comment les coefficients sont pénalisés :
  - Ridge régression : pénalisations des carrées (diapositive précédente)
  - LASSO régression : pénalisations des valeurs absolues :

$$J(w,t) = \sum_{i=1}^{n}(f(X_i) - y_i)^2 \quad \text{subject to} \quad \sum_{j=1}^{d}|w^j| \leq t$$