Data Science

Marc Tommasi

October 3, 2023

Outline

Outline

Main ideas

- Used to explain as well as to predict. Example: explain fuel consumption as a function of car weight.
- Take the class of straight lines $y = ax + b$. Find a and b given weights (x) and observed consumption (y).
- We also call b *intercept* (ordonnée à l'origine) and a is a coefficient (which defines the slope of the line).
- If we know that the data x are centered at 0, then the intercept is zero.
- Generalization to the case where data are represented in a space of dimension d. Example: predict fuel consumption as a function of car weight and engine power, weather, speed, etc.
- We take a class $\mathcal{H}$ of linear functions: we look for w_i such that $y = w_0 + w_1 x_1 + \cdots + w_d x_d$.
- Note: we can pose $x_0 = 1$ and compute

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_d x_d = \boldsymbol{w}^\top \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

.

Formal setting

- $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$
- $\mathcal{H} = \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x} \rangle \mid \boldsymbol{w} \in \mathbb{R}^d\}$ is the class of hypothesis.
- There are m examples of the form $((x_1, x_2, \ldots, x_d), y) = (\boldsymbol{x}, y)$:

$$S = \{(\boldsymbol{x}^{(1)}, y_1), (\boldsymbol{x}^{(2)}, y_2), \ldots, (\boldsymbol{x}^{(m)}, y_m)\}$$

- S can be represented by a matrix X where each row is an instance of S
- We use the quadratic loss (squared error) to measure the empirical error:

$$\ell(h, (\boldsymbol{x}, y)) = (h(\boldsymbol{x}) - y)^2$$

ERM and the objective function

- The aim is to minimize the squared error for each example, and the loss function, the objective function, is the Mean squared error (MSE)

$$\frac{1}{m}\sum_i(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)}\rangle - y_i)^2 = \frac{1}{m}\|X\boldsymbol{w} - \boldsymbol{y}\|^2$$

- The ERM principle is to find the parameters $\boldsymbol{w}$ that minimize the objective function on S:

$$\operatorname*{argmin}_w \frac{1}{m}\|X\boldsymbol{w} - \boldsymbol{y}\|^2$$

- The objective is convex: there is a global minimum and it can be calculated analytically.

Derivatives and gradient

- The derivative f' of a function $f : \mathbb{R} \to \mathbb{R}$ evaluated at x is defined by
$$f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

.

- The derivative cancels for an x that minimizes f.
- For $f : \mathbb{R}^d \to \mathbb{R}$, we can consider the derivative with respect to each component of $\boldsymbol{x} = (x_1, \ldots, x_d)$. This is denoted $\frac{\partial f}{\partial x_i}$.
- The d -dimensional vector of all these derivative functions is the gradient $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})$.
- Remember: the derivative of $f(x) = x^n$ is $f'(x) = nx^{n-1}$ and the derivative of a constant is zero;

Conclusion

- We can find an analytical solution which will be $(X^\top X)^{-1}X^\top \boldsymbol{y}$.
- Informally, we're looking for $X\boldsymbol{w} = \boldsymbol{y}$, which we can rewrite as $X^\top X\boldsymbol{w} = X^\top \boldsymbol{y}$, then $\boldsymbol{w} = (X^\top X)^{-1}X^\top \boldsymbol{y}$.
- Note
  - The matrix $(X^\top X)$ may not be invertible, so we can instead calculate $X^\dagger \boldsymbol{y}$ where $X^\dagger$ is the Moore-Penrose pseudo-inverse, which is calculated by an SVD (If $X = U\Sigma V^\top$ then $X^\dagger = V\Sigma^\dagger U^\top$ and $\Sigma^\dagger$ is calculated by taking the inverse of the values except for the 0s, which remain 0).

Complexity

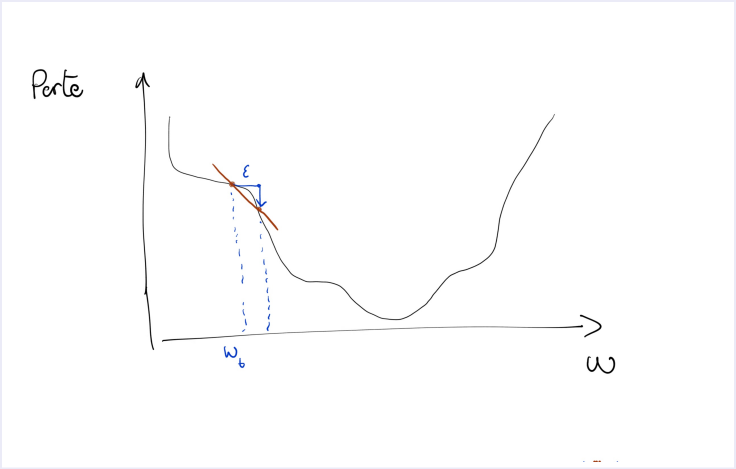
- The matrix inversion is costly for very large datasets
- Matrix multiplication is $O(n^3)$ for simple algorithms ($O(n^{2.4})$ for the best ones).
- The gradient descent approach yields a quadratic algorithm that works quite well in practice.
- (and can be accelerated with GPUs. . .)

Gradient descent I

- The aim is to reduce the empirical error step by step: find the parameter modification that will make the loss function take on a smaller value.
- Idea of descending through the steepest slope on the curve of loss as a function of parameters $\boldsymbol{w}$

Gradient descent II

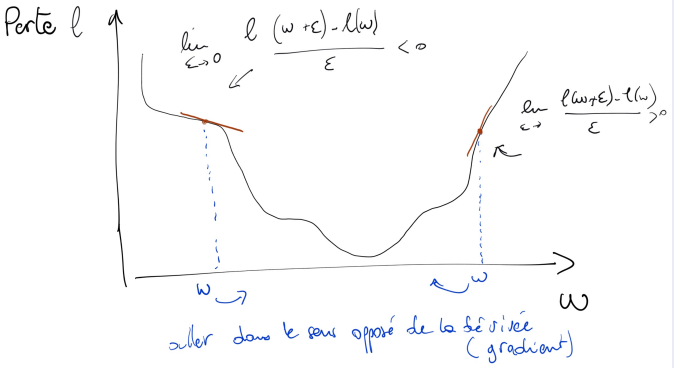
Derivative direction



Quick reminder about derivatives

Gradient descent III

Direction

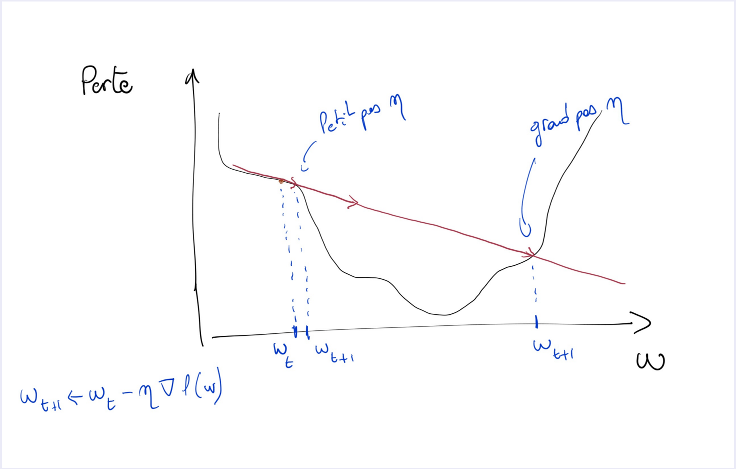


Perte ℓ
$\lim_{\varepsilon \to 0} \frac{\ell(w+\varepsilon)-\ell(w)}{\varepsilon} < 0$
$\lim_{\varepsilon \to} \frac{\ell(w+\varepsilon)-\ell(w)}{\varepsilon} > 0$
w
w
w
aller dans le sens opposé de la dérivée (gradient)

The gradient gives the direction: opposite to the gradient

Gradient descent IV

Step by step



We're taking a step in this direction, by a factor yet to be specified: the learning rate.

Algorithm

- Start with random weights, which define a point on the curve of the objective function.
- Calculate the derivative (gradient) of the objective function at this point
- One step is taken in the opposite direction to this gradient. The length of this step is determined by the learning rate.
- A termination check is performed on the norm of the gradient (tolerance) or on the error.
- Notes
  - In the convex case (as with MSE), a global minimum is guaranteed.
  - It is preferable to have normalized dimensions (use `StandardScaler` for example).
  - Dynamic gradient step can be used: the learning rate value decreases slowly (on the order of $1/t$).

More formally

- Input: Sample $S = \{(\boldsymbol{x}^{(1)}, y_1), (\boldsymbol{x}^{(2)}, y_2), \ldots, (\boldsymbol{x}^{(m)}, y_m)\}$, a loss function f, a learning rate η
- Init: $\boldsymbol{w}^{(0)}$ recieves random values
- Repeat until convergence
  - $t = t + 1$
  - $\boldsymbol{w}^{(t)} \leftarrow \boldsymbol{w}^{(t-1)} - \eta \nabla_{\boldsymbol{w}^{(t-1)}} f(\boldsymbol{w}^{(t-1)}, S)$
- Output : $\boldsymbol{w}^{(t)}$

Batch and stochastic

Gradient descent (Batch Gradient Descent)

- The algorithm requires each iteration to pass over the entire data set (reminder: the loss function is a sum over all examples).
- The cost of this calculation can still be prohibitive
- Sometimes difficult to escape from local minima

Stochastic Gradient Descent (SGD)

- A single example is selected at each round
- Advantages
  - fast
  - if the loss function is very unstable, SGD can get out of local minima

Mini Batch Gradient Descent

- a subset of the examples is selected
- suitable for GPUs, as we can now perform small matrix operations on a batch of data

Polynomial regression

- Linear models have been constructed (straight lines, hyperplanes, etc.).
- This class of hypotheses may be ill-suited to non-linear data.
- We can get around this problem by adding new attributes to the data: square, cube, pairwise products of attributes, etc.
- In the 1-dimension case, let's solve

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_k x^k$$

.

- The data are transformed, but the same linear regression algorithm is applied!
- The same applies if x is a vector in $\mathbb{R}^d$ and products can be made between different attributes ($x_i x_j$, $x_i x_j x_k \ldots$).
- NB: we can avoid materializing the construction of all these new attributes using kernel techniques (used in SVMs, ...).

Technically

- with `PolynomialFeatures`.
- we can use `Pipeline` class objects to compose different stages
  - feature transformation

    ```
    estimators = [('reduce_dim', PCA()), ('clf', SVC())]
    pipe = Pipeline(estimators)
    # two _ to access params
    pipe.set_params(clf__C=10)
    ```

  - application of a model with `pipe.fit`, (`score`, `predict`,. . .)

Error decomposition

Reminder: Complexity/variance bias

- bias: assumption about the function class, for example. Subject to underfitting if bias is too strong
- variance/complexity: variation in function class due to sensitivity to data. Subject to overfitting if the variance is too high.
- NB: there always remains an irreducible error due to data quality in the error decomposition.

Error decomposition

- $L_{\mathcal{D}}(h_S)$ is the sum of the approximation error + the estimation error
- Approximation error
  - Does not depend on S but on the choice of $\mathcal{H}$.
  - Decreases with increasing complexity of $\mathcal{H}$.
- Estimation error
  - decreases if we have more examples in S (we reduce the variance. . .)

Illustration of polynomial regression

- the technique of adding numerous attributes increases complexity and variance.
- the degree of the p polynomials, x_i, x_i^2, $\ldots x_i^p$; the sizes of the products $x_i x_j$, $x_i x_j x_k$, $\ldots$ are hyper-parameters to be adjusted to control bias/complexity.
- we must try to control this:
  - tracing learning curves,
  - regularization

Methodology

- Hyper-parameters must be set on the learning sample.

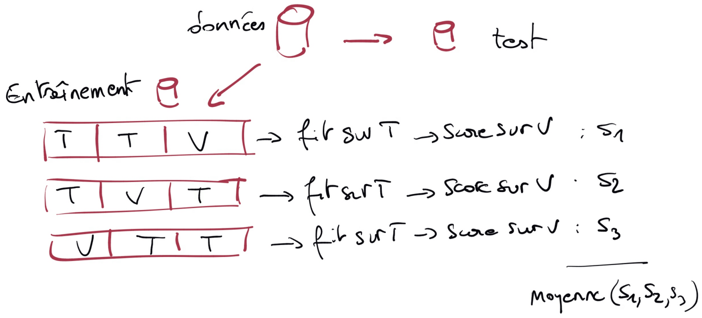
Remember

- The test set should never be looked at, except to evaluate the true/generalization error.

Model selection techniques

- Training set split into learning and validation or cross-validation
- cross-validation,
- leave-one-out,. . .
- `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection`

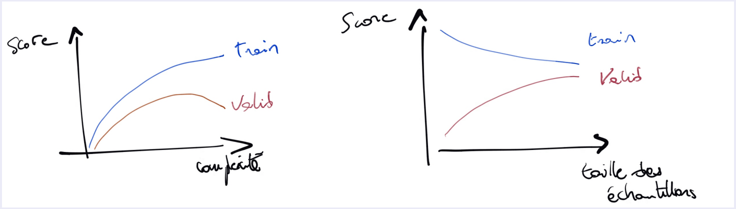
Cross-validation


données
test
Entraînement
T | T | V → fit sur T → score sur V : S_1
T | V | T → fit sur T → score sur V : S_2
V | T | T → fit sur T → score sur V : S_3
Moyenne (S_1, S_2, S_3)

In practice

Learning curves

- calculate the error in training and validation according to sample size or model complexity.
- examining the curve can help select the right parameters



Grid Search

- vary the set of hyper-parameters and perform an exhaustive search for the best set of parameters

Regularization I

Ridge Regression

- Tikonov regularization: adds the square of the norm of the parameter vector ($+\alpha\frac{1}{2}\sum_i w_i^2$). This is also known as a ℓ_2 penalty.
- It's best to normalize the data before applying this regularization.
- Increases the bias and reduces the variance
- A closed form can be calculated, but SGD can also be applied.
- In sklearn, class `Ridge` or `SGDRegressor` with `penalty` which is `"l2"` with parameter `alpha`.

Regularization II

Lasso

- This time it's the ℓ_1 standard, the sum of absolute values ($\alpha \sum_i |w_i|$).
- Removes the least relevant attributes: a sparse model is obtained by variable selection.
- The function is not derivable in 0 (right derivative different from left derivative).
- We can use a sub-differential (sub-gradient, subgradient) around 0, taking the sign if w is non-zero and 0 otherwise.
- Class `Lasso` with parameter `alpha`.

Elastic net

- this is the convex sum of the two regularizations (a ratio is applied between the 2).
- Class `ElasticNet` with parameters `ratio` and `alpha`.

Regularization III

Early stopping

- Limits the number of iterations. This is a form of “algorithmic” regularization.
- We stop when the validation error rises.
- Not easy to identify, as the curves are not smooth!
- parameter `earlystopping` in SGD-type algos.
- can also be simulated with the `warm_start` parameter, which allows optimization to continue where it left off (with `max_iter` to set the number of iterations).
- `SGDRegressor` with zero regularization takes `penalty=None`.