Momentum：结合当前梯度与上一次更新信息。

指数加权平均：$V_t = \beta \cdot V_{t-1} + (1-\beta) \cdot \theta_t$

β：权重　V_{t-1}：过往均值　θ_t：当前参数值

$$= \sum_{i=1}^{N} (1-\beta) \cdot \beta^i \cdot \theta_{N-i}$$

β 表示对 $\frac{1}{1-\beta}$ 次数据的记忆

不考虑 Momentum(m) 时的梯度下降

$W_{i+1} = W_i - lr \cdot g(W_i)$

考虑 m 时(振荡收敛)

$V_i = m \cdot V_{i-1} + g(W_i)$　　V_i：更新量

$W_{i+1} = W_i - lr \cdot V_i$

optim.SGD

params：参数组.

lr：初始学习率

momentum：动量

weight_decay：L2正则化系数.

nesterov：是否采用NAG

其它优化器

optim.Adagrad 自适应学习率梯度下降法
RMSprop：Adagrad的改进
Adadelta：Adagrad的改进
Adam：RMSprop结合Momentum
Adamax：Adam增加学习率上限.
SparseAdam：稀疏版Adam
ASGD：随机平均梯度下降.
Rprop：弹性反向传播
LBFGS：BFGS的改进