CMPUT 267 Basics of Machine Learning

Winter 2024

Outline

1. Recap
2. Gradient Descent

Recap: Estimation

- True data distirubtion is p_{true}.
- Observe dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ where X_i has distribution p_{true}.
- Goal: Estimate properties of p_{true}.

1. Pick a likelihood $p(\mathcal{D} \mid w)$ (data generation model).
   - Pick a distribution class.
2. Pick a prior that is a conjugate prior $p(w)$.
3. Define objective to get w
   - MLE: $c(w) = \ln p(\mathcal{D}|w)$
   - MAP: $c(w) = \ln p(w|\mathcal{D})$
   - Bayesian: $p(w|\mathcal{D})$

Recap: Estimation, Conditional models

- Observe dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of observations x_i and their corresponding targets y_i.
- Goal: Estimate properties of $P_{\text{true}}(Y|X = x)$.
- Our models can be from parametrized families of conditional distributions

$$\mathcal{F} = \{f(y \mid x; \theta) \mid \theta \in \mathbb{R}^k\}$$

1. Pick a likelihood $p(y \mid x, w)$ (data generation model).
   - Pick a distribution class, prior.
   - MLE: $c(w) = \sum_{i=1}^n \ln p(y_i \mid x_i, w)$

Example:

$p(y|x) = N(x w_1, \exp(x w_2))$ $\quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$

MLE

$$\min_{\vec{w} \in \mathbb{R}^2} \sum_{i=1}^{n} -\ln p(y_i | x_i, \vec{w}) \leftarrow \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-y_i)^2}{2\sigma^2}\right)$$

$$-\ln p(y_i|x_i,\vec{w}) = -\ln \frac{1}{\sqrt{2\pi \exp(x_i w_2)}} - \ln\left(\exp\left(-\frac{(x_i w_1 - y_i)^2}{2\exp(x_i w_2)}\right)\right)$$

$$= \ln\left(\sqrt{2\pi}\right) + \frac{1}{2} x_i w_2 + \frac{(x_i w_1 - y_i)^2}{2\exp(x_i w_2)}$$

$$c(w) = \sum_{i=1}^{n} c_i(w) \quad \text{where} \quad c_i(w) = \frac{1}{2}\left[x_i w_2 + \frac{(x_i w_1 - y_i)^2}{\exp(x_i w_2)} \right]$$

$$\frac{\partial c_i(w)}{\partial w_1} = \frac{(x_i w_1 - y_i) x_i}{\exp(x_i w_2)} \qquad \frac{\partial c_i(w)}{\partial w_2} = \frac{x_i}{2} + \frac{(x_i w_1 - y_i)^2}{2} \frac{\partial}{\partial w_2} \exp(-x_i w_2)$$

$$- \exp(-x_i w_2) x_i$$

$$\begin{bmatrix} \frac{\partial c(w)}{\partial w_1} \\ \frac{\partial c(w)}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \frac{\partial c(w)}{\partial w_1} = \sum_{i=1}^{n} \frac{x_i (x_i w_1 - y_i)}{\exp(x_i w_2)} = 0$$

$$\frac{\partial c(w)}{\partial w_2} = \frac{1}{2} \sum_{i=1}^{n} \left[x_i - x_i (x_i w_1 - y_i) \exp(-x_i w_2) \right] = 0$$

→ No closed form ; we need to use gradient descent

GD

Initialize $w_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$t=0$ $\quad g_0 = \nabla c(w_0) = \begin{bmatrix} \frac{\partial c(\vec{w}_0)}{\partial w_1} \\ \frac{\partial c(\vec{w}_0)}{\partial w_2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} -x_i y_i \\ \sum_{i=1}^{n} x_i + x_i y_i \end{bmatrix} \begin{matrix} \leftarrow g_{0,1} \\ \leftarrow g_{0,2} \end{matrix}$

$w_{t+1} \leftarrow w_t - \eta_t g_t$

stopping condition: until $\|g_t\|_2^2 = \sum_{j=1}^{2} g_{t,j}^2 \leq \varepsilon$ threshold

Problem: GD can be very expensive for large dataset

Stochastic Gradient Descent : take a stochastic approximation to $\nabla c(w)$

$$\nabla c(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla c_i(w)$$

→ take a subsample of population

sample average

$\nabla c_K(w)$: K: random sample

$= \frac{1}{|K|}\sum_{i \in K} c_i(w)$

unbiased estimate of $\nabla c(w)$

sample size of 1

$$E[\nabla c_K(w)] = \sum_{i=1}^{n} p(i)\,\nabla c_i(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla c_i(w) = \nabla c(w)$$

$p(i) = 1/n$

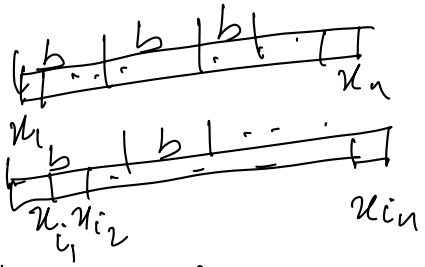
unbiased

But variance may be high.

We might be able to reduce the variance by selecting a batch of samples instead of a single one at each iteration.

SGD with minibatches

batch size of $b > 1$



shuffle →

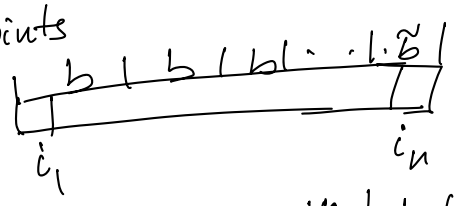
$$\nabla c(w_t) \approx \underbrace{\frac{1}{b}\sum_{j=1}^{b} \nabla c_{ij}(w_t)}_{g_t}$$

$$w_{t+1} = w_t - \eta_t g_t$$

1. Initialize $\vec{w}_0$

For $t = 1$ to T (T: ub of epochs)

shuffle the order of datapoints



For $k = 0, \cdots \left\lfloor \frac{n}{b} \right\rfloor$ $\left\lfloor \frac{n}{b} \right\rfloor$ batches of size b

select k^{th} minibatch

compute g_t from this minibatch

$$g_t = \frac{1}{b} \sum_{i=0}^{b-1} \nabla c_i(w_t)$$

update the parameter guess

$$\vec{w}_{t+1} = \vec{w}_t - \eta_t \vec{g}_t$$

$t = t+1$

Computational cost $w \in \mathbb{R}^d$

GD: $w_{t+1} = w_t - \eta_t \underbrace{\nabla C(w_t)}_{= \frac{1}{n} \sum_{i=1}^{n} \nabla c_i(w_t)}$

cost of one update $O(nd)$

Full cost $O(T_{GD}\, n \cdot d)$

T_{GD}: #updates for GD

SGD $w_{t+1} = w_t - \eta_t \frac{1}{b} \sum_{k=1}^{b} \nabla c_{i_k}(w_t)$

$O(bd)$

$b < n$

$b \approx 32, 64$

$O(T_{SGD} \cdot bd)$

T_{SGD}: # updates for SGD

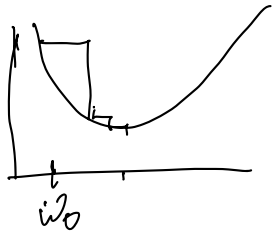
$T_{SGD}\, bd < T_{GD}\, nd \rightarrow T_{SGD} < T_{GD} \cdot \frac{n}{b}$

Eg. n: 1M $b = 32$ $T_{SGD} < 30{,}000\, T_{GD}$

Selection of step size for SGD

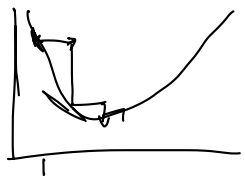
- Fixed step size
  - crude, may take too many iterations to get to optimal point.

- Heuristics $\eta \approx \frac{1}{t}$



- Gradient dependent

$$\eta_t = \frac{1}{1+|g_t|}$$



$$w_{t+1} = w_t - \frac{g_t}{1+|g_t|}$$

$$C(w) = \frac{1}{2}w^2 \quad \nabla C(w) = w$$

If GD:

$w_0 = 1.0 \quad \eta \approx 1/t$

$\eta = \frac{1/2}{t+1}$

$w_1 = w_0 - \eta_0 w_0$

$= 1 - \frac{1}{2}(1)$

$= 1/2$

$\eta_t = \frac{1/2}{2} = 1/4 \quad w_2 = w_1 - \eta_1 w_1$

$= \frac{1}{2} - \frac{1}{4}\left(\frac{1}{2}\right)$

$= 0.375$

$\eta_2 = 0.16\bar{6} \quad w_3 = 0.3125$

$\eta_3 = 0.125 \quad w_4 = 0.273$

$\eta_4 = 0.1 \quad w_5 = 0.246$

Gradient-dep.

$\eta_t = \frac{1}{1+|g_t|}$

$w_0 = 1 \quad \eta_0 = \frac{1}{2}$

$w_1 = w_0 - \eta_0 w_0$

$= 1/2$

$\eta_1 = \frac{1}{1+|g_1|} = \frac{2}{3}$

$w_2 = \frac{1}{2} - \frac{2}{3}\left(\frac{1}{2}\right)$

0.16

$\eta_2 = 0.85 \quad w_3 = 0.02$

$\eta_3 = 0.98 \quad w_4 \approx 5\times10^{-4}$

$\eta_4 = 0.99 \quad w_5 \approx 3\times10^{-7}$

SGD : noisy gradient $\tilde{\nabla} c(w_t) = w_t + \varepsilon_t$

$= \varepsilon_t \sim N(0, 0.2)$

$E[\tilde{\nabla} c(w_t)] = \nabla c(w_t)$

$\eta_t = \frac{1/2}{t+1}$

$w_0 = 1.0$

$w_1 = 0.58$

$w_2 = 0.44$

$w_3 = 0.35$

$w_4 = 0.0.09$

$\vdots$

$w_{1000} \sim 0.003$

$\eta_t = 1/1+|g_t|$

$w_0 = 1.0$

$w_1 = 0.47$

$w_2 = 0.09$

$w_3 = -0.06$

$w_4 = -0.24$

$w_5 = 0.103$

$\vdots$

$w_{10000} \sim 0.05$

Adagrad $\eta_t = \frac{1}{1+\bar{g}_t}$ $\bar{g}_t = \bar{g}_{t-1} + |g_t|$ $\bar{g}_0 = |g_0|$

$\bar{g}_{t+1} \geq \bar{g}_t$ $\eta_{t+1} \leq \eta_t$

slowly decaying