机器学习理论导引
作业五

DZ1833019, 欧先飞,ouxianfei@smail.nju.edu.cn

2019 年 6 月 2 日

1 [20pts] Online Regression

[在线回归] 考虑在线岭回归:

- 每一时刻 t，学习器选择分类面 $\mathbf{w}_t \in \{\mathbf{w} | \|\mathbf{w}\| \leq D\} \subseteq \mathbb{R}^d$;
- 学习器观测到样本和标记 $(\mathbf{x}_t, y_t)$，并遭受损失

$$f_t(\mathbf{w}_t) = (y_t - \mathbf{w}_t^\top \mathbf{x}_t)^2 + \frac{\lambda}{2}\|\mathbf{w}_t\|^2$$

其中 $\lambda > 0$ 是给定的正则化参数，$\|\mathbf{x}_t\| \leq R$，$|y_t| \leq DR$.

请问学习器应该采用什么算法更新分类面 $\mathbf{w}_t$? 采用该算法之后，学习器的遗憾是多少?
要求：描述算法时，需要给出梯度的具体计算公式、如何设置步长。

Analysis.
给定 $(\mathbf{x}_t, y_t)$，则 $\frac{\partial f_t(\mathbf{w}_t)}{\mathbf{w}_t} = \frac{\partial}{\partial \mathbf{w}_t}\left[(y_t - \mathbf{w}_t^{\mathrm{T}}\mathbf{x}_t)^2 + \frac{\lambda}{2}\|\mathbf{w}_t\|_2^2\right] = 2(\mathbf{w}_t^T\mathbf{x} - y_t)\mathbf{x}_t + \lambda\mathbf{w}_t$，由题意有 $\|\mathbf{w}_t\| \leq D$，$\|\mathbf{x}_t\| \leq R$, $|y_t| \leq DR$。

所以有 $\|\frac{\partial f_t(\mathbf{w}_t)}{\mathbf{w}_t}\| \leq 2|\mathbf{w}_t^{\mathrm{T}}\mathbf{x}_t - y_t|\|\mathbf{x}_t\| + \lambda\|\mathbf{w}_t\| \leq 2(DR + DR)R + \lambda D = 4DR^2 + \lambda D$。*参考 pdf 第 32 页定理 3.3，将学学习率设置为* $\eta_t = \frac{2R}{4DR^2 + \lambda D}\sqrt{t}$，*则由定理的结论，可以获得算法的遗憾界为：*

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathbb{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) \leq \frac{3}{2} 2R[4DR^2 + \lambda D]\sqrt{T} = 3DR[4R^2 + \lambda]\sqrt{T} = O(\sqrt{T})$$

而相应的算法为：

Algorithm 1 Online Regression

1: randomly initialize $\mathbf{w}_1 \in \mathcal{W}$
2: **for** t = 1...T **do**
3: $\quad \mathbf{w}'_{t+1} = \mathbf{w}_t - \frac{2R}{(4DR^2 + \lambda D)\sqrt{t}}\left[2(\mathbf{w}_t^{\mathrm{T}}\mathbf{x}_t - y_t)\mathbf{x}_t + \lambda\mathbf{w}_t\right]$
4: $\quad \mathbf{w}_{t+1} = \frac{D}{\max\{D, \|\mathbf{w}'_{t+1}\|\}}\mathbf{w}'_{t+1}$
5: **end for**

2 [25pts] Logistic Regression

[在线分类] 考虑有约束的逻辑回归问题：

- 每一时刻 t，学习器选择分类面 $\mathbf{w}_t \in \{\mathbf{w} | \|\mathbf{w}\| \le D\} \subseteq \mathbb{R}^d$;
- 学习器观测到样本和标记 $(\mathbf{x}_t, y_t)$，并遭受损失

$$f_t(\mathbf{w}_t) = \log\left(1 + \exp(-y_t \mathbf{w}_t^\top \mathbf{x}_t)\right)$$

其中 $\|\mathbf{x}_t\| \le R$，$y_t \in \{+1, -1\}$.

请问学习器应该采用什么算法更新分类面 $\mathbf{w}_t$？采用该算法之后，学习器的遗憾是多少？
要求：描述算法时，需要给出梯度的具体计算公式、如何设置步长。

Analysis.

给定 $(\mathbf{x}_t, y_t)$，则有 $\frac{\partial f_t(\mathbf{w}_t)}{\partial \mathbf{w}_t} = \frac{\partial}{\partial \mathbf{w}_t}[\ln(1+e^{-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t})] = \frac{e^{-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t}}{1+e^{-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t}}(-y_t \mathbf{x}_t)$，进而可以获得其梯度的上界：$\|\frac{\partial f_t(\mathbf{w}_t)}{\partial \mathbf{w}_t}\| \le \|-y_t \mathbf{x}_t\| \le R$。参考定理 *3.3* 可以知道，将步长设置为 $\eta_t = \frac{2R}{R\sqrt{t}} = \frac{2}{\sqrt{t}}$，可以获得目标算法如下：

Algorithm 2 Logistic Regression

randomly initialize $\mathbf{w}_1 \in \mathcal{W}$
for t = 1...T **do**
  $\mathbf{w}'_{t+1} = \mathbf{w}_t - \frac{2}{\sqrt{t}} \frac{e^{-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t}}{1+e^{-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t}}(-y_t \mathbf{x}_t)$
  $\mathbf{w}_{t+1} = \frac{D}{\max\{D, \|\mathbf{w}'_{t+1}\|\}} \mathbf{w}'_{t+1}$
end for

由 *pdf* 第 *32* 页定理 *3.3* 的结论，我们可以获得如下的遗憾界：

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathbb{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) \le \frac{3}{2} 2R \times R\sqrt{T} = 3R^2\sqrt{T} = O(\sqrt{T})$$

由于 $\log(1+\exp(-y_t \mathbf{w}_t^{\mathrm{T}} \mathbf{x}_t))$ 是 α 指数凹的，所以采用在线牛顿法可以获得更好的渐进界，参考 *pdf* 第 *35* 页定理 *3.5* 的证明，使用算法如下：

Algorithm 3 ONS

randomly initialize $\mathbf{w}_1 \in \mathcal{W}$
$A_0 = \frac{4}{\beta^2 D^2} I$
for t = 1...T **do**
  $A_t = A_{t-1} + \nabla f_t(\mathbf{w}_t) \nabla f_t(\mathbf{w}_t)^{\mathrm{T}}$
  $\mathbf{w}'_{t+1} = \mathbf{w}_t - \frac{1}{\beta} A_t^{-1} \nabla f_t(\mathbf{w}_t)$
  $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \mathcal{W}} (\mathbf{w} - \mathbf{w}'_{t+1})^{\mathrm{T}} A_t (\mathbf{w} - \mathbf{w}'_{t+1})$
end for

其中 β 是输入参数，按照定理 *3.5* 的证明，在线牛顿法可以获得 $\frac{1}{2\beta}(1+d\log T) = O(\log T)$ 的渐进界。

3 [20pts] Exponentially Concave

对于二次可微函数 $f(\cdot):\mathcal{W}\mapsto\mathbb{R}$，它是 α 指数凹的充要条件是什么?
提示：利用函数 f 的梯度和海森矩阵建立充要条件。

Analysis.

若函数 $f(\mathbf{w})$ 是 α 指数凹的，则意味着 $\exp(-\alpha f(\mathbf{w}))$ 是凹函数，而这等价于 $\exp(-\alpha f(\mathbf{w}))$ 的海森矩阵半负定。

将 $\exp(-\alpha f(\mathbf{w}))$ 的海森矩阵记为 H，并记 $\mathbf{w}=(w_1,w_2,\ldots,w_d)^{\mathrm{T}}$，那么有：

$$
\begin{aligned}
H_{ij} &= \frac{\partial^2\exp(-\alpha f(\mathbf{w}))}{\partial w_i\partial w_j}\\
&= \frac{\partial}{\partial w_i}\left[\frac{\partial\exp(-\alpha f(\mathbf{w}))}{\partial w_j}\right]\\
&= \frac{\partial}{\partial w_i}\left[\exp(-\alpha f(\mathbf{w}))(-\alpha\frac{\partial f(\mathbf{w})}{\partial w_j})\right]\\
&= \exp(-\alpha f(\mathbf{w}))(-\alpha\frac{\partial f(\mathbf{w})}{\partial w_i})(-\alpha\frac{\partial f(\mathbf{w})}{\partial w_j})+\exp(-\alpha f(\mathbf{w}))(-\alpha)\frac{\partial^2 f(\mathbf{w})}{\partial w_i w_j}\\
&= \alpha^2\exp(-\alpha f(\mathbf{w}))\frac{\partial f(\mathbf{w})}{\partial w_i}\frac{\partial f(\mathbf{w})}{\partial w_j}-\alpha\exp(-\alpha f(\mathbf{w}))\frac{\partial^2 f(\mathbf{w})}{\partial w_i\partial w_j}
\end{aligned}
$$

所以 $H=\alpha^2\exp(-\alpha f(\mathbf{w}))\nabla f(\mathbf{w})\nabla f(\mathbf{w})^{\mathbf{T}}-\alpha\exp(-\alpha f(\mathbf{w}))\nabla^2 f(\mathbf{w})$，由定义知：$f$ 是 α 指数凹的等价于 H 半负定。

4 [25pts] Online-to-Batch Conversation

对于在线凸优化问题, 假设 $f_1,\ldots,f_T$ 是从同一分布 $\mathcal{P}$ 独立采样得到。假设随机函数 $f_1,\ldots,f_T$ 为凸，其定义域 $\mathcal{W}$ 直径小于 D，梯度的范数小于 G，根据讲义介绍，在线梯度下降有如下遗憾界:

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}) \leq \frac{3DG}{2}\sqrt{T} \tag{4.1}$$

令 $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$，$F(\cdot) = \mathbb{E}_{f\sim\mathcal{P}}[f(\cdot)]$。在遗憾界 (4.1) 的基础上，证明以大概率

$$F(\bar{\mathbf{w}}) - F(\mathbf{w}) = O\left(\frac{1}{\sqrt{T}}\right).$$

提示：假设随机函数 f 有界，然后利用讲义定理 2.2 (针对鞅的 Azuma 不等式)

Proof.

首先对 $F(\bar{\mathbf{w}}) - F(\mathbf{w})$ 进行变换：

$$\begin{aligned}
F(\bar{\mathbf{w}}) - F(\mathbf{w}) &= F(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t) - F(\mathbf{w}) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) - F(\mathbf{w}) \\
&= \frac{1}{T}\sum_{t=1}^{T} [F(\mathbf{w}_t) - F(\mathbf{w})] \\
&= \frac{1}{T}\sum_{t=1}^{T} [F(\mathbf{w}_t) - f_t(\mathbf{w}_t) + f_t(\mathbf{w}_t) - f_t(\mathbf{w}) + f_t(\mathbf{w}) - F(\mathbf{w})] \\
&= \frac{1}{T}\sum_{t=1}^{T} [F(\mathbf{w}_t) - f_t(\mathbf{w}_t)] + \frac{1}{T}\sum_{t=1}^{T} [f_t(\mathbf{w}) - F(\mathbf{w})] + \frac{1}{T}\sum_{t=1}^{T} [f_t(\mathbf{w}_t) - f_t(\mathbf{w})]
\end{aligned}$$

由题目中*4.1*知，上述不等式右边最后一项的上界为 $\frac{3DG}{2\sqrt{T}}$，而由于 $\mathbb{E}[f_t(\mathbf{w}_t)] = F(\mathbf{w}_t)$、$\mathbb{E}[f_t(\mathbf{w})] = F(\mathbf{w})$，所以上述不等式右边前两项，从 $t = 1$ 到 T 构成分别构成两个鞅差序列。假设 $|f_t(\cdot)| \leq c$，则对鞅差序列中每一项有 $|F(\mathbf{w}_t) - f_t(\mathbf{w}_t)| \leq 2c$，进而由 *Azuma* 不等式可以得到：$\Pr\left[\sum_{t=1}^{T}(F(\mathbf{w}_t) - f_t(\mathbf{w}_t)) \geq \epsilon\right] \leq \exp(-\frac{\epsilon^2}{8Tc^2})$，令 $\exp(-\frac{\epsilon^2}{8Tc^2}) = \delta$ 可以得到：

$$\Pr\left[\sum_{t=1}^{T}(F(\mathbf{w}_t) - f_t(\mathbf{w}_t)) \geq \sqrt{8Tc^2\ln\frac{1}{\delta}}\right] \leq \delta$$

对于不等式右边第二项同理可得 $\Pr\left[\sum_{t=1}^{T}(f_t(\mathbf{w}_t - F(\mathbf{w}_t))) \geq \sqrt{8Tc^2\ln\frac{1}{\delta}}\right] \leq \delta$。将这两个结论代入上述不等式,则可以得到 $\Pr\left[F(\bar{\mathbf{w}}) - F(\mathbf{w}) \leq \frac{2}{T}\sqrt{8Tc^2\ln\frac{1}{\delta}} + \frac{3DG}{2\sqrt{T}}\right] \geq 1-\delta$,而 $\frac{2}{T}\sqrt{8Tc^2\ln\frac{1}{\delta}} + \frac{3DG}{2\sqrt{T}} = 4c\sqrt{\frac{\ln 1/\delta}{T}} + \frac{3DG}{2\sqrt{T}} = O(\frac{1}{\sqrt{T}})$。

□

5 [10pts] Comparison with SGD

在讲义定理 2.1 中，我们证明了 SGD 同样可以达到

$$F(\bar{\mathbf{w}}) - F(\mathbf{w}) = O\left(\frac{1}{\sqrt{T}}\right).$$

请问 SGD 的证明过程和习题 4 的证明过程有什么异同?

Analysis.
(1) 两者都利用凸函数的条件对当前解与最优解的上界进行评估，并利用等式 $\langle \mathbf{a}-\mathbf{b}, \mathbf{a}-\mathbf{c}\rangle = \frac{1}{2}(\|\mathbf{a}-\mathbf{b}\|_2^2 + \|\mathbf{a}-\mathbf{c}\|_2^2 - \|\mathbf{b}-\mathbf{c}\|_2^2)$ 将上界转化为逐差的形式，然后对 T 项累加以放缩上界。
(2) 不同的地方在于 *SGD* 的证明过程直接对函数空间 $\mathcal{F}$ 的期望 F 进行放缩，而习题 *4* 中的证明中所利用到的遗憾界是从采样得到的函数 f_t 出发开始放缩。