Statistical Machine Learning
Homework3

方言
2021210929

2022 年 5 月 16 日

1 Learning Theory

1.1 Problem 1

1.1.1 Please show that the regression with squared error can be also viewed as a special case of the general setting.

对于回归任务，给定输入输出的集合 X, Y，我们有:

$$\begin{aligned}
&Z = X \times Y \\
&H = \{f : f \text{ is a mapping from } X \text{ to } Y\} \\
&\ell(h, (x, y)) = (h(x) - y)^2
\end{aligned}$$

1.2 Problem 2

1.2.1 Prove that the generalization error $R(h, D)$ is always $\frac{1}{2}$.

由于在二分类任务下，对 $\forall h \in H$，即对 $\forall h$, h 是一个 X 到 $\{0, 1\}$ 的映射:

$$\begin{aligned}
R(h, D) &= \mathbb{E}_{z \sim D}[\ell(h, z)] \\
&= P(\{h(x) = 0\}) \cdot \mathbb{E}_{h(x)=0}[\ell(h, z)] + P(\{h(x) = 1\}) \cdot \mathbb{E}_{h(x)=1}[\ell(h, z)] \\
&= P(\{h(x) = 0\}) \cdot \mathbb{E}_{h(x)=0}[1_{y \neq 0}] + P(\{h(x) = 1\}) \cdot \mathbb{E}_{h(x)=1}[1_{y \neq 1}] \\
&= P(\{h(x) = 0\}) \cdot \frac{1}{2} + P(\{h(x) = 1\}) \cdot \frac{1}{2} \\
&= \frac{1}{2} \cdot (P(\{h(x) = 0\}) + P(\{h(x) = 1\})) \\
&= \frac{1}{2}
\end{aligned}$$

1.3 Problem 3

1.3.1 Let D be a data distribution and $R^\star$ be the infimum of the generalization error on D, prove the following inequality.

由于 $h_S^{ERM} \in \underset{h \in H}{\arg\min}\, \ell(h, S)$，即:

$$h_S^{ERM} \in \underset{h \in H}{\arg\min}\, \hat{R}(h, S)$$

因此:

$$\hat{R}(h_S^{ERM}, S) - \hat{R}(h, S) \leq 0, \quad \forall h \in H$$

令 $h^\star \in \arg\min_{h \in H} R(h, D)$，则:

$$R^\star - R(h, D) \leq R(h^\star, D) - R(h, D) \leq 0, \quad \forall h \in H$$

其中，$R(h^\star, D)$ 可以无限逼近 $R^\star$，即可认为二者相等。

由此可得:

$$\begin{aligned} R(h_S^{ERM}, D) - R^\star &\leq \left(R(h_S^{ERM}, D) - R^\star\right) - \left(\hat{R}(h_S^{ERM}, S) - \hat{R}(h^\star, S)\right) \\ &= \left(R(h_S^{ERM}, D) - R(h^\star, D)\right) - \left(\hat{R}(h_S^{ERM}, S) - \hat{R}(h^\star, S)\right) \\ &= \left(R(h_S^{ERM}, D) - \hat{R}(h_S^{ERM}, S)\right) - \left(R(h^\star, D) - \hat{R}(h^\star, S)\right) \\ &\leq 2 \sup_{h \in H} |R(h, D) - \hat{R}(h, S)| \end{aligned}$$

1.3.2 Give a generalization error bound of ERM.

由于 H 是有限的，因此给定分布 D 和 $S \sim D^m$，对 $\forall h \in H$:

$$R(h, D) = \mathbb{E}[\ell(h, z)], \quad \hat{R}(h, S) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

又因为 $\ell(h, z_i)$ 是相互独立的，且 $\ell(h, z_i) \in [0, M]$，则根据 Hoeffding 不等式，有:

$$P_{S \sim D^m} \left(|R(h, D) - \hat{R}(h, S)| > \frac{\epsilon}{2} \right) \leq 2 \exp(-\frac{m\epsilon^2}{2M^2})$$

又由上题已知:

$$R(h_S^{ERM}, D) - R^\star \leq 2 \sup_{h \in H} |R(h, D) - \hat{R}(h, S)|$$

由此可得，对于 $\forall \epsilon > 0$:

$$\begin{aligned} P_{S \sim D^m} \left(R(h_S^{ERM}, D) - R^\star > \epsilon \right) &\leq P_{S \sim D^m} \left(2 \sup_{h \in H} |R(h, D) - \hat{R}(h, S)| > \epsilon \right) \\ &= P_{S \sim D^m} \left(\sup_{h \in H} |R(h, D) - \hat{R}(h, S)| > \frac{\epsilon}{2} \right) \\ &= P_{S \sim D^m} \left(\exists h \in H: \ |R(h, D) - \hat{R}(h, S)| > \frac{\epsilon}{2} \right) \\ &\leq \sum_{h \in H} P_{S \sim D^m} \left(|R(h, D) - \hat{R}(h, S)| > \frac{\epsilon}{2} \right) \\ &\leq 2|H| \exp(-\frac{m\epsilon^2}{2M^2}) \end{aligned}$$

令 $\delta = 2|H| \exp(-\frac{m\epsilon^2}{2M^2})$，可以解出:

$$\epsilon = \sqrt{\frac{2M^2(\ln 2|H| + \ln \delta^{-1})}{m}}$$

由于:

$$P_{S \sim D^m} \left(R(h_S^{ERM}, D) - R^\star > \epsilon \right) \leq \delta$$

因此:

$$\begin{aligned} &P_{S \sim D^m} \left(R(h_S^{ERM}, D) - R^\star \leq \epsilon \right) \\ =&1 - P_{S \sim D^m} \left(R(h_S^{ERM}, D) - R^\star > \epsilon \right) \\ \geq&1 - \delta \end{aligned}$$

即对于 $\forall\delta\in(0,1)$:

$$P_{S\sim D^m}\left(R(h_S^{ERM},D)-R^\star>\sqrt{\frac{2M^2(\ln 2|H|+\ln\delta^{-1})}{m}}\right)\geq 1-\delta$$

2 Dimension Reduction, PCA

2.1 Problem 4

2.1.1 Choose d to preserve different information and show the resulting images.

对于保留 $1\%, 5\%, 20\%, 50\%, 80\%, 95\%, 99\%$ 的信息，分别计算可得，对应的 d 为 $1, 1, 3, 11, 44, 154, 331$。

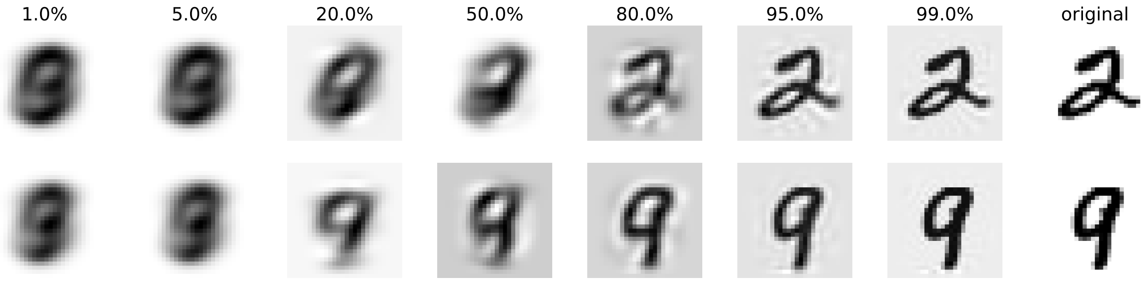

图 1: 保留不同信息下的 PCA 结果

2.1.2 Visualize the top 100 eigen vectors to see how they look

对 top100 的特征向量使用 t-SNE 降到 2 维进行可视化，可以看到其分布比较均匀。

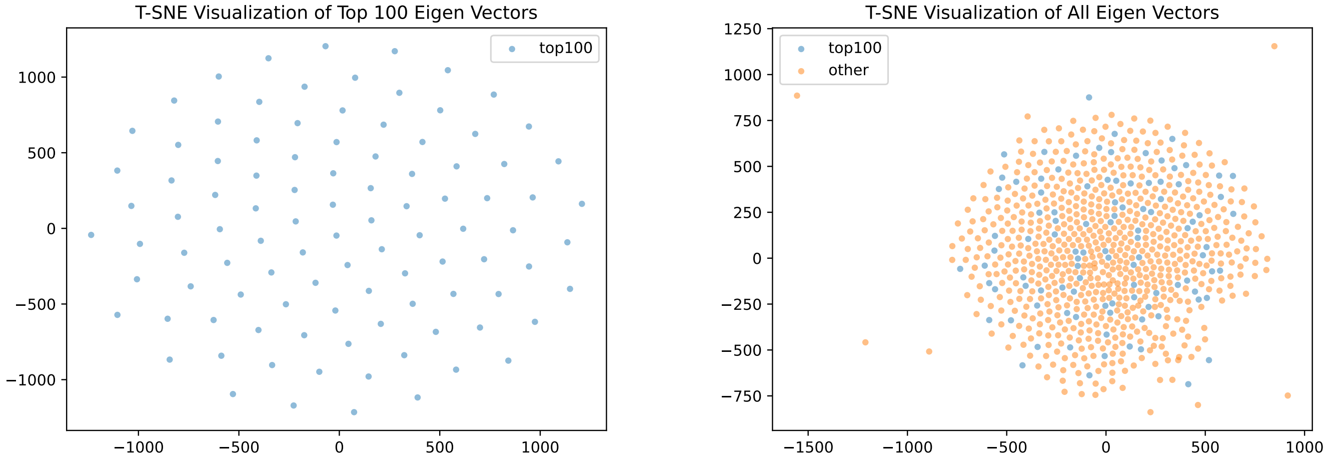

图 2: t-SNE 可视化，Top100 特征向量 (左) / 所有特征向量 (右)

2.1.3 Re-run PCA without centering the dataset.

没有进行中心化操作的情况下，对于保留 $1\%, 5\%, 20\%, 50\%, 80\%, 95\%, 99\%$ 的信息，分别计算可得，对应的 d 为 $1, 1, 1, 3, 23, 103, 281$。

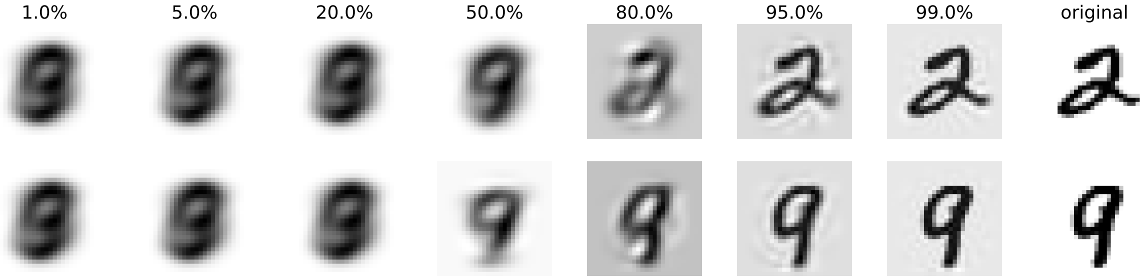

图 3: 保留不同信息下的 PCA 结果，无中心化

对比可以看出，在保留比较少的信息时候 (5%~80%)，中心化操作会提升 PCA 重建图像的质量，即 PCA 损失的信息更少。在要求保留比较多的信息时，二者没有明显区别。