arXiv:2303.02658v1 [cs.LG] 5 Mar 2023

关于特权经验风险最小化的容量界限

Michal Sharoni
内盖夫本-古里安大学计算机科学系，以色列贝尔谢巴

Sivan Sabato
内盖夫本-古里安大学计算机科学系，以色列贝尔谢巴

摘要

我们研究一种称为利用特权信息学习的监督学习范式，该范式最早由 Vapnik 和 Vashist（2009年）提出。在这一范式中，除了样本和标签之外，仅针对训练样本提供额外的（特权）信息。其目标是利用这些信息提高所得分类器的分类准确率；对于新的样本实例，该分类器只能使用其中的非特权信息来预测其标签。我们在 Pechyony 和 Vapnik（2010a）提出的自然特权经验风险最小化算法下，研究带零一损失的特权学习理论。我们针对该工作中关于这一问题所诱导的损失类 VC维 的一项论断给出了一个反例；我们得出结论，该论断是不正确的。随后，我们对 VC维 进行了正确的分析，给出了特权经验风险最小化损失类容量的下界和上界。我们进一步通过泛化分析表明，除非特权信息的容量与非特权信息的容量相近或更小，否则特权经验风险最小化的最坏情况保证不可能优于标准的非特权经验风险最小化。该结果指出了特权经验风险最小化方法的一项重要局限。在结尾的讨论中，我们提出了另一种可能性：即使特权信息的容量很大，特权经验风险最小化仍可能发挥作用。

1 引言

经典的监督式机器学习范式考虑如下情形：给定一组带标签的训练样本，尝试从给定的函数集合中找出一个在未知测试样本上泛化误差最小的函数。在这项工作中，我们研究了对这一设定的一种扩展。该扩展最初由 Vapnik 和 Vashist（2009年）提出，称为利用特权信息学习，简称特权学习。在这一范式中，训练阶段会向学习者提供有关训练样本的额外信息。这类信息称为特权信息，仅在训练阶段对训练样本可用。目标是利用这些信息提高最终分类器的分类准确率。分类器本身只能利用新样本的非特权信息来预测其标签。因此，特权信息只有在有助于获得更好的分类器时才有用。

这一范式的一个经典启发性示例（见 Vapnik和 Vashist，2009年）考虑了这样一种情形：目标是根据手术前可获得的患者信息，找到一条能够预测手术三个月后结果的规则。然而，对于以往的患者，还收集了手术期间及手术后的额外信息。虽然对新患者进行分类时无法获得这些信息，但历史数据中确实存在这些信息，因此可在训练期间将其用作特权信息。

在这项工作中，我们研究了 Pechyony 和 Vapnik（2010a）提出的一种自然的经验风险最小化算法，称为特权经验风险最小化。该算法最小化非特权信息与特权信息的联合损失。我们给出了新的结果，指出当特权信息是高维的，或者更一般地说，当相关的特权损失类具有较高容量时，这种方法存在局限性。在许多场景中，高维特权信息是很自然的：用于训练的离线测量具有比测试时的测量更高的带宽或灵敏度。例如，考虑一个学习问题，其目标是对图像进行分类，

其中，非特权信息提供低分辨率图像，而特权信息提供高分辨率图像。如果训练样本在训练期间可以使用先进设备进行扫描，而分类器部署在实地的低资源环境中，只能获取低质量图像，那么就会出现这种情况。Lee 等人（2020）研究了类似的应用。我们在此表明，在这一情形下，采用零一损失的特权经验风险最小化方法若无额外假设，便无法保证取得成功。

我们对特权经验风险最小化算法所诱导的损失类进行了 VC维分析。我们的分析包括一个反例，反驳了 Pechyony 和 Vapnik（2010a）此前提出的一项论断；该论断中的错误可追溯至其证明中的一处疏漏。我们给出了正确的分析，并提供了 VC维的下界和上界。随后，我们研究了在何种情况下可以保证特权经验风险最小化比完全不使用特权信息的标准经验风险最小化获得更优的误差界。我们得出结论：此类最坏情况保证必须以特权信息类具有较低容量为前提。最后，我们提出了一种可能的方式，说明即使特权信息的容量较大，特权经验风险最小化仍可能发挥作用。

2 相关工作

特权学习这一范式最早由 Vapnik 和 Vashist（2009 年）提出。这项工作引入了 SVM+ 算法，展示了如何通过改变 SVM 类算法的目标，使其纳入特权信息。除了引入 SVM+ 算法之外，Vapnik 和 Vashist（2009年）还推导出相关结果，表明当特权信息类的维度较低时，在这类算法中利用特权信息可以提高收敛速率。Pechyony 和 Vapnik（2010a）提出了一种称为特权经验风险最小化的经验风险最小化算法，并将特权学习中的优化问题推广到其他损失函数。他们还针对该算法的收敛速率提出了若干理论论断。

自诞生以来，特权学习已应用于各种领域。Lapin 等人（2014）研究了支持向量机+ 与加权支持向量机之间的联系。他们证明，特权信息可以通过与每个训练样本关联的权重进行编码。此外，他们还证明，加权支持向量机总能复现支持向量机+ 的解，而反之则不成立。Vapnik 和 Izmailov（2015）描述了两种与实例空间和特权信息空间之间知识迁移相关的机制。这些机制可用于加快学习速度。Qi 等人（2015）提出了一种利用特权信息的半监督学习方法。该方法能够利用未标记数据中的分布信息和特权信息，提高学习效率。Yang 等人（2016）在经验风险最小化框架下提出了一种度量学习算法，该算法利用特权信息放宽了先前的度量学习方法。Vrigkas 等人（2016）描述了一种结合利用特权信息学习与主动学习的概率方法。Pasunuri 等人（2016）提出了一种利用特权信息学习决策树的算法。Vapnik 和 Izmailov（2017）提出了一种将知识从特权信息空间迁移到特征空间的机制。他们证明，该机制既适用于神经网络框架，也适用于支持向量机。近期研究探讨了视觉领域中的特权学习（例如，Yuan 等人，2019年；Gao 等人，2019年；Li 等人，2019年）。Lee 等人（2020）研究了一项特权信息具有高维度的应用。然而，特权学习理论尚未探讨其基本方法下特权信息的容量上限。

3 预备知识与设定

我们首先描述 Pechyony 和 Vapnik（2010a）定义的一般损失下的特权学习设定。令 $\mathcal{X}$ 为我们希望进行标注的元素定义域。令 $\mathcal{X}^*$ 为训练样本中可用的特权信息的定义域。令 $\mathcal{Y}$为可能标签的集合。学习器的输入由一系列独立同分布的三元组构成：

$$S = (x_1, x_1^*, y_1), ..., (x_m, x_m^*, y_m),$$
$$x_i \in \mathcal{X}, \quad x_i^* \in \mathcal{X}^*, \quad y_i \in \mathcal{Y}, \tag{1}$$

这些三元组根据固定但未知的概率分布 $\mathcal{D}$ 生成，该分布定义在 $\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y}$上。令 $\ell_{\mathcal{X}}$: $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ 为非特权样本定义域上的有界损失函数。特权学习的目标是利用包含特权信息的样本 S，找到一个在 $\mathcal{D}$上损失较低的假设。

假设特权信息的损失有界，$\ell_{\mathcal{X}^*}$:$\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$。令 $C > 0$ 为一个常数，并记 $[t]_+ = \max(t, 0)$。给定一个仅使用非特权信息的分类器 h: $\mathcal{X} \to \mathcal{Y}$，以及一个特权信息函数 ϕ: $\mathcal{X}^* \to \mathcal{Y}$，Pechyony 和 Vapnik（2010a）将复合假设 (h, ϕ) 在样本 (x, x^*, y) 上的损失定义为：

$$\ell'_C(h, \phi, (x, x^*, y)) =$$
$$\frac{1}{C}\ell_{\mathcal{X}^*}(\phi(x^*), y) + [\ell_{\mathcal{X}}(h(x), y) - \ell_{\mathcal{X}^*}(\phi(x^*), y)]_+.$$

函数 ϕ 被视为对示例中由 h 引起的损失进行“校正的函数”。给定一个函数

类，该函数类基于非特权信息 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$，以及一个基于特权信息 $\Phi \subseteq \mathcal{Y}^{\mathcal{X}^*}$的函数类，Pechyony 和 Vapnik（2010a）将特权经验风险最小化问题定义为如下优化问题：

$$\min_{h\in\mathcal{H},\phi\in\Phi} \sum_{i=1}^{m} \ell'_C((h,\phi),(x_i,x_i^*,y_i)). \tag{2}$$

在这项工作中，我们关注上述设定中的一个重要特例：二元标签（$\mathcal{Y}=\{0,1\}$）和二元损失函数，且 $C=1$。在这种情况下，我们有

$$\ell'_C(h,\phi,(x,x^*,y)) = \max\{\ell_{\mathcal{X}}(h(x),y), \ell_{\mathcal{X}^*}(\phi(x^*),y)\}.$$

由此得到以下优化问题：

$$\min_{h\in\mathcal{H},\phi\in\Phi} \sum_{i=1}^{m} \max\{\ell_{\mathcal{X}}(h(x_i),y_i), \ell_{\mathcal{X}^*}(\phi(x_i^*),y_i)\}. \tag{3}$$

可以将 ϕ 视为指示在对 h上的损失进行最小化时应考虑哪些训练样本，其中 $\phi(x_i^*) = 1$ 表示在最小化过程中应忽略样本 x_i。例如，如果特权信息能够识别标记的可靠性，那么这就可能很有用，例如众包标签的情形。因此，我们假设 $\ell_{\mathcal{X}}$ 是关于非特权信息的标准损失，定义如下：

$$\ell^{01}_{\mathcal{X}}(\hat{y},y) := \mathbf{1}[\hat{y}\neq y]$$

并且 $\ell_{\mathcal{X}^*}$ 是关于特权信息的“忽略”损失，定义如下：

$$\ell^{\text{ig}}_{\mathcal{X}^*}(z,y) := \mathbf{1}[z=1].$$

将 h 相对于分布 $\mathcal{D}$ 的误差记为$\text{err}(h,\mathcal{D}) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}[h(X)\neq Y] = \mathbb{E}[\ell_{\mathcal{X}}(h(X),Y)]$。令$\text{err}(h,S)$表示 h 在 S上的均匀分布下的经验误差。因此，特权学习的目标是从 $\mathcal{H}$ 中找到一个假设，使其在 $\mathcal{D}$上的误差较低，并使用样本 S。在我们此处研究的特权经验风险最小化范式中，这是通过求解式（3）中的优化问题来实现的。

4 VC 维度分析

本节研究式（3）中定义的最小化问题所涉及函数类的 VC 维度。用 VC(·) 表示一个函数类的 VC 维度。令 $d := \text{VC}(\mathcal{H})$，$d^* := \text{VC}(\Phi)$。定义

$$f_{(h,\phi)}((x,x^*),y) := \max(\ell^{01}_{\mathcal{X}}(h(x),y), \ell^{\text{ig}}_{\mathcal{X}^*}(\phi(x^*),y)),$$

并令

$$\mathcal{F}_{(\mathcal{H},\Phi)} = \{f_{(h,\phi)} \mid h\in\mathcal{H}, \phi\in\Phi\}.$$

当下标可由上下文明确判断时，我们用 $\mathcal{F}$ 表示 $\mathcal{F}_{(\mathcal{H},\Phi)}$。公式 (3) 等价于在 S上使用假设类 $\mathcal{F}$运行经验风险最小化。因此，特权经验风险最小化的泛化行为由 $\mathcal{F}$的 VC维刻画。

VC($\mathcal{F}$) 与 VC($\mathcal{H}$)、VC(Φ) 的值之间有什么关系？Pechyony 和 Vapnik（2010a）似乎已经回答了这个问题；他们定义了以下损失类：

$$\mathcal{L}(\mathcal{H}) := \{\ell_{\mathcal{X}}(h(\cdot),\cdot) \mid h\in\mathcal{H}\},$$
$$\mathcal{L}(\Phi) := \{\ell_{\mathcal{X}^*}(\phi(\cdot),\cdot) \mid \phi\in\Phi\},$$
$$\mathcal{L}(\mathcal{H},\Phi) = \{\ell'_C((h,\phi),(\cdot,\cdot,\cdot)) \mid h\in\mathcal{H},\phi\in\Phi\}.$$

并声称以下等式成立：[1]

Claim of Pechyony and Vapnik (2010a):
$$\text{VC}(\mathcal{L}(\mathcal{H},\Phi)) = \text{VC}(\mathcal{L}(\mathcal{H})) + \text{VC}(\mathcal{L}(\Phi)). \tag{4}$$

随后，研究者利用该等式证明了特权经验风险最小化优化问题的泛化上界。

对于 $\ell_{\mathcal{X}} := \ell^{01}_{\mathcal{X}}$ 和 $\ell_{\mathcal{X}^*} := \ell^{\text{ig}}_{\mathcal{X}^*}$，有 VC($\mathcal{H}$) =VC($\mathcal{L}(\mathcal{H})$)，VC($\Phi$) = VC($\mathcal{L}(\Phi)$)，且 VC($\mathcal{L}(\mathcal{H},\Phi)$) =VC($\mathcal{F}_{(\mathcal{H},\Phi)}$)。因此，若式（4）成立，则可推出 VC($\mathcal{F}$) = VC($\mathcal{H}$) +VC($\Phi$)。然而，我们现在将证明，这个等式实际上并不成立。[2] 下面的定理 4.1 给出了一个反例，反驳了所声称的式（4）。

定理 4.1. 对于任意整数 $d > 0$，以及任意两个定义域 $\mathcal{X}$，$\mathcal{X}^*$，使得 $|\mathcal{X}|,|\mathcal{X}^*| \geq 3d$，存在假设类 $\mathcal{H}_d \subseteq \{0,1\}^{\mathcal{X}}$和 $\Phi_d \subseteq \{0,1\}^{\mathcal{X}^*}$，使得 VC($\mathcal{H}_d$) = VC($\Phi_d$) = d，而 VC($\mathcal{F}_{(\mathcal{H}_d,\Phi_d)}$) = $3d$。

证明. 首先，考虑 $d = 1$的情况。令 $X_3 = \{x_1,x_2,x_3\} \subseteq \mathcal{X}$ 为来自 $\mathcal{X}$的一个包含三个定义域样本的集合。我们通过三元组 ($h(x_1)$ $h(x_2)$ $h(x_3)$)，描述一个从 X_3 到$\{0,1\}$ 的假设 h。类似地，令$X_3^* = \{x_1^*,x_2^*,x_3^*\} \subseteq \mathcal{X}^*$，并通过三元组 ($\phi(x_1^*)$ $\phi(x_2^*)$ $\phi(x_3^*)$)，描述一个从 X_3^*到 $\{0,1\}$ 的假设 ϕ。定义以下关于 X_3 和 X_3^*的假设类：

$$\mathcal{H}_1 := \{(0,0,0),(0,0,1),(1,0,0),(1,1,0)\},$$
$$\Phi_1 := \{(0,0,0),(0,0,1),(0,1,0),(1,0,1)\}.$$

容易看出，VC($\mathcal{H}_1$) = VC(Φ_1) = 1。另一方面，将 $\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$ 限制在集合 $X_3 := \{((x_i,x_i^*),0)\}_{i\in[3]}$ 上，可得 VC($\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$) ≤ 3，因为定义域的大小为 3。此外，VC维恰好为 3，因为 $\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$ 在

[1]原始论断涵盖实值损失，因此需要推广 VC 维度的定义；此处我们仅针对映射到 $\{0,1\}^2$ 的损失这一特殊情况陈述该定义。我们发现，问题源于 Pechyony 和 Vapnik（2010a）中公式 (4) 的证明里量词顺序使用错误；该证明可在完整版中查阅（Pechyony 和 Vapnik，2010b）。

~X_3：任意标记方式 $h \in \mathcal{H}_1$ 都可以通过 f_{h,ϕ_0} 得到，其中ϕ_0 是 Φ_1 中的全零函数。类似地，Φ_1 中的所有标记方式都可以通过使用全零 $h_0 \in \mathcal{H}_1$得到。另外两个缺失的标记方式是 (0, 1, 1) 和 (1, 1, 1)。前者可以通过 $h =$ (0, 0, 1) 和 $\phi =$ (0, 1, 0) 得到，后者可以通过 $h =$ (1, 1, 0) 和$\phi =$ (0, 0, 1) 得到。因此，VC($\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$) $= 3$，如所声称。

接下来，考虑 $d > 1$。令 $X_{3d} = \{x_1, ..., x_{3d}\} \subseteq \mathcal{X}$为由 $\mathcal{X}$ 中的 $3d$ 个不同定义域点组成的集合。将这些点划分为 d 个三元组，记作 $t_1 := (x_1, x_2, x_3)$ ..., $t_d := (x_{3d-2}, x_{3d-1}, x_{3d})$。我们将定义域 X_{3d}上的假设 h 描述为一系列来自 $\mathcal{H}_1$ 的 d 个函数，它们分别应用于三元组 $t_1, \ldots, t_d$中的样本。定义域 X^*_{3d} 上的 ϕ 的描述方式类似。定义以下从 X_{3d} 到 $\{0, 1\}$ 以及从 X^*_{3d} 到 $\{0, 1\}$的函数假设类：

$$\mathcal{H}_d := \{(h_1, ..., h_d) \mid h_1, ..., h_d \in \mathcal{H}_1\},$$
$$\Phi_d := \{(\phi_1, ..., \phi_d) \mid \phi_1, ..., \phi_d \in \Phi_1\}.$$

我们首先证明 VC($\mathcal{H}_d$) $= d$。反设 VC($\mathcal{H}_d$) $> d$。那么存在一个含有$d + 1$ 个点的打散集。由于预测器定义在 d 个三元组上，根据鸽巢原理，打散集中必有两个点来自同一个三元组。由此可得，$\mathcal{H}_1$ 能打散一个大小为二的集合，这与 VC($\mathcal{H}_1$) $= 1$矛盾。因此，VC($\mathcal{H}_d$) $\leq d$。接下来，我们通过证明存在一个大小为 d的打散集，来证明 VC($\mathcal{H}$) $\geq d$。由于 VC($\mathcal{H}_1$) $= 1$，在每个 d 个定义域三元组中，都有一个点被 $\mathcal{H}_1$ 打散。所有这些点构成一个大小为 d 的集合，并且该集合被 $\mathcal{H}_d$打散，证毕。因此，VC($\mathcal{H}_d$) $= d$。通过类似的论证，VC(Φ_d) $= d$。

最后，我们证明 VC($\mathcal{F}_{(\mathcal{H}_d,\Phi_d)}$) $= 3d$。定义集合

$$\mathcal{F}_d := \{(f_1, ..., f_d) \mid f_1, ..., f_d \in \mathcal{F}_{(\mathcal{H}_1,\Phi_1)}\}.$$

我们首先断言 $\mathcal{F}_d \subseteq \mathcal{F}_{(\mathcal{H}_d,\Phi_d)}$：令 ($f_{(h_1,\phi_1)}$ ⋯, $f_{(h_d,\phi_d)}$) $\in \mathcal{F}_d$。根据$\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$的定义，$\forall i \in [d]$ 我们有

$$f_{(h_i,\phi_i)}((x, x^*), y) = \max(\ell^{01}_{\mathcal{X}}(h_i(x), y), \ell^{\text{ig}}_{\mathcal{X}^*}(\phi_i(x^*), y)).$$

因此，

$$(f_{(h_1,\phi_1)}, ..., f_{(h_d,\phi_d)}) = f_{((h_1,...,h_d),(\phi_1,...,\phi_d))} \in \mathcal{F}_{(\mathcal{H}_d,\Phi_d)}.$$

由此我们得出 VC($\mathcal{F}_{(\mathcal{H}_d,\Phi_d)}$) $\geq$ VC($\mathcal{F}_d$)。~现在，考虑集合 $X_{3d} = \{((x_i, x^*_i), 0)\}_{i\in[3d]}$。将 $\mathcal{F}_d$ 限制到集合 X_{3d}，得到在 X_{3d}中每个三元组上的所有可能~函数，这与 $d = 1$的情形相同。~因此，$\mathcal{F}_d$ 被 X_{3d}打散。所以，VC($\mathcal{F}_d$) $\geq 3d$，如需证明。 □

我们在以下定理中给出了 VC($\mathcal{F}_{(\mathcal{H},\Phi)}$) 的正确上界。

定理4.2. 设 d、d^* 为整数，$\mathcal{X}$ 为某个定义域。设 $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$、$\Phi \subseteq \{0, 1\}^{\mathcal{X}^*}$ 为假设类，且 VC($\mathcal{H}$) $= d$，VC(Φ) $= d^*$。则

$$\text{VC}(\mathcal{F}_{(\mathcal{H},\Phi)}) \leq 4 \log_2(4e)(d+d^*+1) \approx 13.77(d+d^*+1).$$

为了证明该定理，我们首先给出同一定义域上两个假设类并集的 VC维的紧上界。

引理 4.3. 设 $\mathcal{Y} = \{0, 1\}$。令 $\mathcal{J}$、$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ 为定义在同一定义域上的两个假设类。则

$$\text{VC}(\mathcal{H} \cup \mathcal{J}) \leq \text{VC}(\mathcal{H}) + \text{VC}(\mathcal{J}) + 1.$$

此界是紧的：对于任意 d、$d^* \in \mathbb{N}$，存在 $\mathcal{J}$ 和 $\mathcal{H}$，使得 VC($\mathcal{H}$) $= d$，VC($\mathcal{J}$) $= d^*$，且 VC($\mathcal{H}\cup\mathcal{J}$) $= d + d^* + 1$.

证明。对于假设类 $\mathcal{F}$，将 $\mathcal{F}$的增长函数记为

$$\Pi_{\mathcal{F}}(m) := \max\{\, |\mathcal{F}|_S| \mid S \subseteq \mathcal{X} \times \mathcal{Y}, |S| = m \,\}.$$

显然，对于任意 S，$|\mathcal{H} \cup \mathcal{J}|_S \leq |\mathcal{H}|_S + |\mathcal{J}|_S$。因此 $\Pi_{\mathcal{H}\cup\mathcal{J}}(m) \leq \Pi_{\mathcal{H}}(m) + \Pi_{\mathcal{J}}(m)$。根据 Sauer引理（Sauer，1972），并利用恒等式 $\binom{m}{k} = \binom{m}{m-k}$，记 $d :=$ VC($\mathcal{H}$) 和 $d^* :=$ VC($\mathcal{J}$)，可得

$$\Pi_{\mathcal{H}\cup\mathcal{J}}(m) \leq \sum_{i=0}^{d}\binom{m}{i} + \sum_{i=0}^{d^*}\binom{m}{i} = \sum_{i=0}^{d}\binom{m}{i} + \sum_{i=m-d^*}^{m}\binom{m}{i}. \tag{5}$$

若 $m > d + d^* + 1$，则 $m - d^* \geq d + 2$，因此：

$$\sum_{i=0}^{d}\binom{m}{i} + \sum_{i=m-d^*}^{m}\binom{m}{i} \leq \sum_{i=0}^{m}\binom{m}{i} = 2^m. \tag{6}$$

结合式 (5) 和式 (6)，我们得出，对于 $m > d + d^* + 1$，$\Pi_{\mathcal{H}\cup\mathcal{J}}(m) < 2^m$。因此，

$$\text{VC}(\mathcal{H} \cup \mathcal{J}) \leq d + d^* + 1.$$

这证明了上界。

为了证明这个界是紧的，设 d，$d^* \in \mathbb{N}$，并考虑一个大小为 $d + d^* + 1$ 的定义域 $\mathcal{X}$。令 $\mathcal{H}$ 为所有将 $\mathcal{X}$ 中至多 d 个样本映射到 1 的函数所构成的集合，令 $\mathcal{J}$ 为所有将 $\mathcal{X}$ 中至多 d^* 个样本映射到 0 的函数所构成的集合。于是，VC($\mathcal{H}$) $= d$，且 VC($\mathcal{J}$) $= d^*$。为了证明 VC($\mathcal{H} \cup \mathcal{J}$) $\geq d + d^* + 1$，考虑对 $\mathcal{X}$ 的标记的两种情况：

- 如果标记中至多包含 d 个正标签，那么 $\mathcal{H}$ 中存在一个函数可以给出该标记。

- 如果标记中包含超过 d 个正标签，那么由于定义域的大小为 $d+d^*+1$，该标记至多包含 d^* 个负标签。因此， $\mathcal{J}$ 中存在一个函数可以给出该标记。

因此， $\mathcal{H}\cup\mathcal{J}=\mathcal{Y}^{\mathcal{X}}$，所以 VC维($\mathcal{H}\cup\mathcal{J}$) $=|\mathcal{X}|=d+d^*+1$，如所述。 □

在定理4.2的证明中，我们还使用了 Blumer 等人（1989）提出的一个定理。[3]

定理 4.4（Blumer 等人，1989）。设（$\mathcal{X},\mathcal{R}$）是一个集合系统，其中 $\mathcal{X}$ 是一个元素集合， $\mathcal{R}$ 是 $\mathcal{X}$的子集集合。对于整数 $k\geq 2$ 和集合族 $\mathcal{R}$，定义$\mathcal{R}^{k\cup}$:
$=\{R_1\cup...\cup R_k \mid R_1,...,R_k\in\mathcal{R}\}$，即 $\mathcal{R}$的 k 重并集。则

$$\mathrm{VC}(\mathcal{R}^{k\cup})\leq \mathrm{VC}(\mathcal{R})\cdot 2k\log_2(2ek).$$

现在我们来证明定理4.2。

定理4.2的证明。设 $\mathcal{H},\Phi$ 为定理中所述的假设类。定义 $\mathcal{H}',\Phi'\subseteq\{0,1\}^{\mathcal{X}\times\mathcal{X}^*\times\{0,1\}}$如下:

$$\mathcal{H}':=\{(x,x^*,y)\mapsto \ell^{01}_{\mathcal{X}}(h(x),y)\mid h\in\mathcal{H}\}$$

and

$$\Phi':=\{(x,x^*,y)\mapsto \ell^{\mathrm{ig}}_{\mathcal{X}^*}(\phi(x^*),y)\mid \phi\in\Phi\}.$$

此外，令

$$\mathcal{F}'_{\mathcal{H},\Phi}:=\{(x,x^*,y)\mapsto f_{(h,\phi)}((x,x^*),y)\mid h\in\mathcal{H},\phi\in\Phi\}.$$

不难看出，VC($\mathcal{H}'$) = VC($\mathcal{H}$)、VC(Φ') =VC(Φ)，以及 VC($\mathcal{F}'_{\mathcal{H},\ \Phi}$) = VC($\mathcal{F}_{\mathcal{H},\ \Phi}$)。此外，对于任意 $h\in\mathcal{H}$、$\phi\in\Phi$，我们有

$$\begin{aligned}\{(x,x^*,y)\mid f_{(h,\phi)}((x,x^*),y)=1\}=\\ \{(x,x^*,y)\mid \ell^{01}_{\mathcal{X}}(h(x),y)=1\}\\ \cup\{(x,x^*,y)\mid \ell^{\mathrm{ig}}_{\mathcal{X}^*}(\phi(x^*),y)=1\}.\end{aligned}$$

因此，将 $\{0,1\}^{\mathcal{X}\times\mathcal{X}^*\times\{0,1\}}$ 中的函数视为集合，我们有 $f_{(h,\phi)}=h'\cup\phi'$，其中 $h'\in\mathcal{H}'$ 和 $\phi'\in\Phi'$。令 $\mathcal{R}:=\mathcal{H}'\cup\Phi'$，则有 $\mathcal{F}_{(\mathcal{H},\Phi)}\subseteq\mathcal{R}^{2\cup}$。因此，VC($\mathcal{F}$) $\leq$VC($\mathcal{R}^{2\cup}$)。根据取 $k=2$的定理 4.4，可得 VC($\mathcal{F}$) $\leq$VC($\mathcal{R}$)$\cdot 4\log_2(4e)$。此外，根据引理 4.3，VC($\mathcal{R}$) =VC($\mathcal{H}\cup\Phi$) $\leq d+d^*+1$。因此，VC($\mathcal{F}$) $\leq 4\log_2(4e)(d+d^*+1)$，如所述。 □

[3]在 Blumer 等人（1989）中，定理陈述并未明确说明对 k 的依赖关系；我们从证明中提取了确切常数。

5 特权经验风险最小化的容量限制

现在我们转而研究式 (3) 中优化问题解的收敛速率，并推导对 VC维取值的条件，使该界能够优于常规经验风险最小化的已知界。我们表明，为了保证泛化改进，特权信息的 VC维不能比非特权信息的 VC维大得太多，这限制了该方法在零一损失情形下的实用性。

当式（3）中优化问题的误差保证小于标准经验风险最小化泛化界的误差保证时，就能保证实现泛化改进。如先前的研究（例如，Vapnik和Vashist，2009）所示，这要求使用能够考虑相对于假设类的误差的界：特权信息的优势（如果存在）来自由于错误率较低而可能实现更快收敛。Boucheron 等人（2005）给出了能够考虑误差的紧误差界。固定 $m\in\mathbb{N}$，对于 $d\in\mathbb{N}$、$x\in[0,1]$，记为

$$\begin{aligned}R_{\mathrm{f}}(d)&:=\frac{8d\log(m+1)+4\log(\frac{4}{\delta})}{m}\\ R_{\mathrm{s}}(x,d)&:=\sqrt{x\cdot R_{\mathrm{f}}(d)},\end{aligned}$$

其中， R_{f} 表示快速速率， R_{s} 表示慢速率。根据 Boucheron 等人（2005，推论5.2），令 h_{ERM} 表示假设类 $\mathcal{H}$ 在样本 S上的标准经验风险最小化算法的输出，其训练误差为 $\hat{}\varepsilon_{\mathrm{ERM}}$:= err(h_{ERM}, S)，则以至少 $1-\delta$的概率，

$$\mathrm{err}(\hat{h}_{\mathrm{ERM}},\mathcal{D})\leq\hat{\varepsilon}_{\mathrm{ERM}}+R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}},d)+R_{\mathrm{f}}(d):=\mathrm{B}_{\mathrm{ERM}}.$$

基于这一结果，我们为特权经验风险最小化的情形推导出一个类似的上界。为给出该上界，我们先定义一些记号。定义一个辅助损失函数

$$\ell^a_{(h,\phi)}(x,x^*,y):=\mathbf{1}[h(x)\neq y\wedge\phi(x^*)=0]$$

以及损失类$\mathcal{L}^a_{(\mathcal{H},\Phi)}:=\{\ell^a_{(h,\phi)}\mid h\in\mathcal{H},\phi\in\Phi\}$。记 d_a:= VC($\mathcal{L}^a_{(\mathcal{H},\Phi)}$)。根据定理4.2， $d_a\leq 13.77(d+d^*+1)$。以下引理给出了d_a的下界，由此得出结论： $d_a=\Theta(d+d^*)$。

引理 5.1. 对于 $\mathcal{H}$、Φ，满足 d、$d^*>1$，

$$d_a\geq d+d^*-2.$$

证明。设 $C_{\mathcal{H}}=\{x_1,...,x_d\}\subseteq\mathcal{X}$ 是一个大小为 d 的集合，它被 $\mathcal{H}$ 打散；设 $C_\Phi=\{x^*_1,...,x^*_{d^*}\}\subseteq\mathcal{X}^*$ 是一个

大小为 d^* 的集合，它被 Φ 打散。定义

$$C_1 := \{(x_i, x^*_{d^*}, 0) \mid 1 \le i \le d-1\},$$
$$C_2 := \{(x_d, x^*_j, 0) \mid 1 \le j \le d^*-1\},$$
$$C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}} := C_1 \cup C_2.$$

注意，$|C_{\mathcal{L}^a(\mathcal{H},\Phi)}| = |C_1| + |C_2| = d + d^* - 2$。现在我们证明，$C_{\mathcal{L}^a(\mathcal{H},\Phi)}$ 被 $\mathcal{L}^a$ 打散。

令 $L = (l_1, \ldots, l_{d+d^*-2}) \in \{0,1\}^{d+d^*-2}$ 为 $\mathcal{L}^a$ 的一个潜在标记。对于每个 $p = (x, x^*_{d^*}, 0) \in C_1$，令 $l(p)$ 为根据 L 对 p 进行的标记。令 $h \in \mathcal{H}$ 满足：对于每个 $p = (x, x^*_{d^*}, 0) \in C_1$，$h(x) = l(p)$，且 $h(x_d) = 1$。这样的 $h \in \mathcal{H}$ 存在，因为 $C_{\mathcal{H}}$ 被 $\mathcal{H}$ 打散。同样，对于每个 $p^* = (x_d, x^*, 0) \in C_2$，令 $l(p^*)$ 为根据 L 对 p^* 进行的标记。令 $\phi \in \Phi$ 满足：对于每个 $p^* = (x_d, x^*, 0) \in C_2$，$\phi(x^*) = 1 - l(p^*)$，且 $\phi(x^*_{d^*}) = 0$。这样的 $\phi \in \Phi$ 存在，因为 C_Φ 被 Φ 打散。现在我们声称，$\ell^a_{(h,\phi)}$ 对于 $C_{\mathcal{L}^a(\mathcal{H},\Phi)}$ 获得标记 L：对于每个 $p = (x, x^*_{d^*}, 0) \in C_1$，

$$\ell^a_{(h,\phi)}(x, x^*_{d^*}, 0) = \mathbf{1}[h(x) \ne 0 \wedge \phi(x^*_{d^*}) = 0]$$
$$= \mathbf{1}[h(x) \ne 0] = l(p).$$

此外，对于每个 $p^* = (x_d, x^*, 0) \in C_2$，

$$\ell^a_{(h,\phi)}(x_d, x^*, 0) = \mathbf{1}[h(x_d) \ne 0 \wedge \phi(x^*) = 0]$$
$$= \mathbf{1}[\phi(x^*) = 0] = l(p^*).$$

我们得出结论：$C_{\mathcal{L}^a(\mathcal{H},\Phi)}$ 被 $\mathcal{L}^a$ 打散。由于 $|C_{\mathcal{L}^a(\mathcal{H},\Phi)}| = d + d^* - 2$，$\mathrm{VC}(\mathcal{L}^a_{(\mathcal{H},\Phi)}) \ge d + d^* - 2$，如所述。 □

$\hat{}\hat{}$用 (h, ϕ) 表示在式 (3) 的优化问题中取得最小值的某个赋值。用下式表示经验忽略权重：

$$\widehat{\varepsilon}_{ig} := \frac{1}{m} \sum_{(x,x^*,y)\in S} \mathbf{1}[\hat{\phi}(x^*) = 1].$$

这是在对 $\mathcal{H}$ 上的误差进行最小化时，因特权信息而被忽略的训练样本所占的比例。用下式表示经验未解释误差：

$$\widehat{\varepsilon}_u := \frac{1}{m} \sum_{(x,x^*,y)\in S} \mathbf{1}[\ell^a_{(\hat{h},\hat{\phi})}(x, x^*, y) = 1].$$

这是未被 ϕ 忽略、但仍被 h 错误分类的训练样本所占的比例。以下定理给出了特权经验风险最小化的泛化误差界。

定理 5.2。以概率 $1 - 2\delta$ 对 $S \sim \mathcal{D}^m$ 的随机选择，

$$\mathrm{err}(\hat{h}, \mathcal{D}) \le \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a)$$
$$+ R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) := \mathrm{B}_{\mathrm{PR}}.$$

证明。设 $\mathcal{D}'$ 是关于 $(\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y})$ 的一个分布，$\times \{0\}$ 使得关于 $(\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y})$ 的边缘分布为 $\mathcal{D}$。回顾一下，$\hat{}\hat{}\hat{}(h, \phi)$ 是公式 (3) 的最小化解。将 h 的误差分解如下：

$$\mathrm{err}(\hat{h}, \mathcal{D}) =$$
$$= \mathbb{P}[\hat{h}(X) \ne Y \wedge \hat{\phi}(X^*) = 1]$$
$$+ \mathbb{P}[\hat{h}(X) \ne Y \wedge \hat{\phi}(X^*) = 0]$$
$$\le \mathbb{P}[\hat{\phi}(X^*) = 1] + \mathbb{P}[\hat{h}(X) \ne Y \wedge \hat{\phi}(X^*) = 0]$$
$$= \mathrm{err}(\hat{\phi}, \mathcal{D}') + \mathrm{err}(\ell^a_{(\hat{h},\hat{\phi})}, \mathcal{D}'), \quad (7)$$

$\hat{}\hat{}$其中，我们将 ϕ 视为等价于 $(x, x^*, y) \mapsto \phi(x^*)$。

我们将分别界定右侧的每一项。

给定 $S = ((x_i, x^*_i, y_i))_{i\in[m]} \sim \mathcal{D}^m$，令 $S_0 := (((x_i, x^*_i, y_i), 0))_{i\in[m]}$，使得 S_0 服从来自 $\mathcal{D}'$ 的独立同分布样本。则 $\mathrm{err}(\ell^a{}_{(\hat{h},\hat{\phi})} S_0) = \widehat{\varepsilon}_u$ 且 $\hat{}\mathrm{err}(\phi, S_0) = \widehat{\varepsilon}_{ig}$。

$\tilde{}$根据 Boucheron 等人（2005），给定一个样本 $S = ((a_1, b_1) \ldots, (a_m, b_m))$，它是根据分布 $\tilde{\mathcal{D}}$ 在 $\mathcal{A} \times \{0,1\}$ 上的生成结果，且假设类 $\mathcal{J} \subseteq \{0,1\}^{\mathcal{A}}$ 的 VC 维为 d'，以至少 $1 - \delta$ 的概率，对于$\tilde{}$所有 $g \in \mathcal{J}$，若 $\hat{}\varepsilon = \mathrm{err}(g, S)$，则

$$\mathrm{err}(g, \tilde{\mathcal{D}}) \le \hat{\varepsilon} + R_{\mathrm{s}}(\hat{\varepsilon}) + R_{\mathrm{f}}(d'). \quad (8)$$

$\tilde{}$ $\tilde{}$令 $\mathcal{D} := \mathcal{D}', \mathcal{J} := \mathcal{L}^a_{\mathcal{H},\Phi}, S := S_0$，可得

$$\mathrm{err}(\ell^a_{(\hat{h},\hat{\phi})}, \mathcal{D}') \le \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d_a). \quad (9)$$

$\tilde{}$ $\tilde{}$此外，令 $\mathcal{D} := \mathcal{D}', \mathcal{F} := \Phi, S := S_0$，代入式 (8)，可得

$$\mathrm{err}(\hat{\phi}, \mathcal{D}') \le \widehat{\varepsilon}_{ig} + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{f}}(d^*). \quad (10)$$

结合式 (9) 和式 (10) 与式 (7)，并使用并集界，以至少 $1 - 2\delta$ 的概率，

$$\mathrm{err}(\hat{h}, \mathcal{D}) \le \widehat{\varepsilon}_{ig} + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{f}}(d^*)$$
$$+ \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d_a),$$

如所声称的那样。 □

定理 5.2 中的上界是利用经验风险最小化在有界无先验误差情形下的已知上界推导得出的。尽管已知这些上界对于零一损失是紧的，但对于我们用于 Φ 预测的损失，并不存在等价的结果。下面的定理表明，尽管如此，经典的无先验一致收敛上界对于这种损失也是紧的。证明见附录 A.1。

定理 5.3。设 Φ 是一个假设类，其 VC 维 $(\Phi) = d^*$。对于所有 $\varepsilon \in (0,1)$ 和 $\delta < 1/128$，如果样本

量为 $m < (d^* - 1)/(1280 \cdot \varepsilon^2)$，则存在一个分布 $\mathcal{D}$，使得以大于 δ的概率，ˆ$\exists \phi \in \Phi$ 满足 $\mathbb{P}[\phi(X) = 1] - \mathbb{P}[\phi(X) = 1] > \varepsilon$，ˆ其中 $\mathbb{P}$ 表示基于大小为 m的随机独立同分布样本计算的经验概率。

我们希望推导出满足 $\mathrm{B}_{\mathrm{PR}} < \mathrm{B}_{\mathrm{ERM}}$ 的条件。然而，这些界分别使用了不同的经验测度。接下来的引理将这两组测度联系起来，说明无论忽略哪一组样本，经验风险最小化算法的经验误差都小于式 (3) 中的最小化值。

引理 5.4. 对任意 $S' \subseteq S$ 和任意 $h \in \mathcal{H}$,

$$m \cdot \hat{\varepsilon}_{\mathrm{ERM}} \leq \sum_{(x_i,y_i) \in S \setminus S'} \mathbf{1}[h(x_i) \neq y_i] + |S'|.$$

由此可得 ˆ$\varepsilon_{\mathrm{ERM}} \leq \widehat{\varepsilon}_u + \widehat{\varepsilon}_{ig}$.

证明. 设 $h \in \mathcal{H}$ 和 $S' \subseteq S$。根据定义，我们有 ˆ$\varepsilon_{\mathrm{ERM}} \leq \mathrm{err}(h, S)$。此外，

$$\begin{aligned} m \cdot \mathrm{err}(h,S) &= \sum_{(x_i,y_i) \in S} \mathbf{1}[h(x_i) \neq y_i] \\ &= \sum_{(x_i,y_i) \in S \setminus S'} \mathbf{1}[h(x_i) \neq y_i] + \sum_{(x_i,y_i) \in S'} \mathbf{1}[h(x_i) \neq y_i] \\ &\leq \sum_{(x_i,y_i) \in S \setminus S'} \mathbf{1}[h(x_i) \neq y_i] + |S'|. \end{aligned}$$

这就证明了该结论。 □

这个引理对于比较 B_{PR} 和 $\mathrm{B}_{\mathrm{ERM}}$ 至关重要，因为它意味着，要得到 $\mathrm{B}_{\mathrm{PR}} < \mathrm{B}_{\mathrm{ERM}}$，唯一的办法就是使 B_{PR} 中的收敛项比 $\mathrm{B}_{\mathrm{ERM}}$ 中的更小。

接下来，我们推导出一个使 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$ 成立的充分条件。该条件考虑的是一种最佳情形：它要求特权信息使经验风险最小化恰好忽略那些无法使用 $\mathcal{H}$ 正确分类的样本。在这种情形下，只要未解释误差足够小，特权学习界就小于经验风险最小化界。该定理的证明见附录 A.2。

定理 5.5。假设 ˆ$\varepsilon_{\mathrm{ERM}} = \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u$。假设$\delta$固定，且 d、d^*、d_a 都很大。那么，如果

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \Theta\left(\frac{\sqrt{d} - \sqrt{d^*}}{\sqrt{d_a}}\right) - \sqrt{\frac{\log(m)}{m}} \cdot \Theta\left(\frac{d_a + d^* - d}{\sqrt{d_a}}\right),$$

那么 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$。

注意，根据引理 5.1，第二项必然为正。

当经验误差较小时，上述充分条件更加严格。特别是在可实现情形下，即 ˆ$\varepsilon_{\mathrm{ERM}} = 0$，该充分条件永远不成立。此外，只有在 $d^* \leq d$ 时（并且样本量较小），该充分条件才可能成立。因此，这并不允许特权类别 Φ 具有较大的容量。事实上，下面的结果表明，一般而言，d^* 不可能远大于 d，同时仍允许 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$。

定理 5.6。对于任意固定的 $\delta \in (0,1)$，如果 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$，那么

$$d^* \leq 2.25 \cdot d + o(1).$$

最后一项的收敛是相对于 d、d^* 共同增长和/或 m 的增长而言的。

该定理在附录 A.3 中得到证明。

6 讨论

我们的工作表明，在零一损失的情况下，用于特权学习的特权经验风险最小化方法在特权信息类上存在固有的容量限制。当经验风险最小化能够被准确执行时，这一分析与特权学习的统计收益相关。然而，当使用替代损失时，情况可能大不相同。在这些情况下，特权学习可能具有计算优势，因为可以利用特权信息使最小化真实损失的计算问题变得更容易。例如，如果特权信息能够识别异常值，从而帮助忽略某些训练样本，并以此使替代损失的优化目标更接近真实目标损失，那么最终的训练误差可能会降低，进而导致更低的真实误差。我们计划在未来的工作中研究这一富有前景的方向。

我们进一步指出，我们的分析仅在最坏情况分析下、且在特定的特权经验风险最小化框架内，对特权信息的维度给出了限制。研究该框架的其他变体，或许能得出限制较少的结果。

参考文献

Martin Anthony 和 Peter L Bartlett。《神经网络学习：理论基础》。剑桥大学出版社，2009年。

Anselm Blumer、Andrzej Ehrenfeucht、David Haussler 和 Manfred K Warmuth。《可学习性与 Vapnik–Chervonenkis 维度》。美国计算机协会会刊（JACM），36(4):929–965，1989年。St´ephane Boucheron、Olivier Bousquet 和 G´abor Lugosi。《分类理论：近期一些进展综述》。ESAIM：概率与统计，9:323–375，2005年。

Zhifan Gao、Sitong Wu、Zhi Liu、Jianwen Luo、Heye Zhang、Mingming Gong 和 Shuo Li。《利用特权信息学习超声弹性成像中的隐式应变重建》。医学图像分析，58:101534，2019年。

Maksim Lapin、Matthias Hein 和 Bernt Schiele。利用特权信息学习：支持向量机+ 与加权支持向量机。《神经网络》，53：95–108，2014年。

Wonkyung Lee、Junghyup Lee、Dohyung Kim 和 Bum-sub Ham。利用特权信息学习以实现高效图像超分辨率。载于《欧洲计算机视觉会议》，第465–482页。施普林格，2020年。

Xun Yang、Meng Wang、Luming Zhang 和 Dacheng Tao。利用特权信息进行度量学习的经验风险最小化。载于 IJCAI，第 2266–2272 页，2016年。

Shanxin Yuan、Bjorn Stenger 和 Tae-Kyun Kim。利用深度数据通过特权学习从 RGB 图像进行三维手部姿态估计。载于 IEEE/CVF 国际计算机视觉会议研讨会论文集，第 0–0 页，2019年。

Yan Li、Fanqing Meng 和 Jun Shi。利用特权信息学习可改善阿尔茨海默病的基于神经影像的计算机辅助诊断：一项比较研究。《医学与生物工程与计算》，57(7)：1605–1616，2019年。Rahul Pasunuri、Phillip Odom、Tushar Khot、Kristian Kersting 和 Sriraam Natarajan。利用特权信息学习：决策树与提升。载于国际人工智能联合会议（IJCAI）研讨会论文集，2016年。Dmitry Pechyony 和 Vladimir Vapnik。关于利用特权信息学习的理论。载于《神经信息处理系统进展》，第1894–1902页，2010a。Dmitry Pechyony 和 Vladimir Vapnik。关于利用特权信息学习的理论（完整版），2010年1月。网址：https://www.researchgate.net/publication/228565062_On_the_Theory_of_Learning_with_Privi

特权

Zhiquan Qi、Yingjie Tian、Lingfeng Niu 和 Bo Wang。利用特权信息进行半监督分类。国际机器学习与控制论杂志，6(4):667–676，2015年。

Norbert Sauer。关于集合族的密度。组合理论杂志A辑，13(1):145–147，1972年。Vladimir Vapnik 和 Rauf Izmailov。利用特权信息学习：相似性控制与知识迁移。机器学习研究杂志，16(2023-2049):2，2015年。Vladimir Vapnik 和 Rauf Izmailov。支持向量机和神经网络中的知识迁移。数学与人工智能年刊，81(1-2):3–19，2017年。Vladimir Vapnik 和 Akshay Vashist。一种新的学习范式：利用特权信息学习。神经网络，22(5-6):544–557，2009年。

Michalis Vrigkas、Christophoros Nikou 和 Ioannis A Kakadiaris。从弱标注样本中主动特权学习人类活动。载于2016 IEEE国际图像处理会议（ICIP），第3036–3040页。IEEE，2016年。

A 延后证明

下面给出上述定理的延后证明。附录 A.1 给出了下界定理 5.3 的证明，附录 A.2 给出了充分条件定理 5.5 的证明，附录 A.3 给出了必要条件定理 5.6 的证明。

A.1 特权学习损失下界的证明

定理 5.3 的证明。该证明改编自 Anthony 和 Bartlett（2009年，定理 5.2）中针对零一损失下界的证明，并将其应用于我们的设定。主要难点在于构造一组分布，使得只有使用最坏情况下数量的样本才能将它们区分开来。以下新构造实现了这一点。

由于 Φ 的 VC 维度为 d^*，因此存在一个包含 d^* 个样本的集合 $C = \{x_1^*, ..., x_{d^*}^*\}$，该集合被 Φ 打散。为简单起见，假设 d^* 是偶数。如果 d^* 是奇数，则下面的证明对 $d^* - 1$ 仍然成立。我们将集合 C 划分为 $d^*/2$ 对 $\{(a_i, b_i)\}_{i \in [d^*/2]}$。考虑定义在 $\mathcal{X} \times \mathcal{Y}$ 上的所有分布 $\mathcal{D}$，它们满足以下性质：

- $\mathcal{D}$ 对所有与 $C \times \{0, 1\}$ 不相交的集合赋予零概率。
- 对于 $x \in \mathcal{X}$，记 $\mathcal{D}(x) = \mathbb{P}_{(X,\ Y) \sim \mathcal{D}}[X = x]$。令 $\alpha := \frac{8\varepsilon}{(1-8\delta)}$。对于每个 $i = 1$、2、...、$d^*/2$，以及 C 的划分中的一对 (a_i, b_i)，以下情况之一成立：
  - $\mathcal{D}(a_i) = \frac{1+\alpha}{d^*}$ 且 $\mathcal{D}(b_i) = \frac{1-\alpha}{d^*}$，或者
  - $\mathcal{D}(b_i) = \frac{1+\alpha}{d^*}$ 和 $\mathcal{D}(a_i) = \frac{1-\alpha}{d^*}$。

令 $\Phi' \subseteq \Phi$ 为包含所有假设 ϕ 的集合，其中对于每一对 (a_i, b_i)，ϕ 将该对中的一个元素映射为 0，另一个映射为 1。

给定 $\mathcal{D}$，令 $\phi^* \in \Phi'$ 为满足以下条件的函数：对于每一对 (a_i, b_i)，当且仅当 $\mathcal{D}(a_i) = \frac{1-\alpha}{d^*}$ 时，$\phi^*(a_i) = 1$。于是，

$$\mathbb{P}[\phi^*(X) = 1] = \sum_{i=1}^{d^*/2} \frac{1-\alpha}{d^*} = \frac{1-\alpha}{2}.$$

此外，对于任意 $\phi \in \Phi'$，我们有

$$\mathbb{P}[\phi(X) = 1] = \sum_{i=1}^{d^*/2} \left(\frac{1+\alpha}{d^*} \mathbf{1}[\phi(a_i) \neq \phi^*(a_i)] + \frac{1-\alpha}{d^*} \mathbf{1}[\phi(a_i) = \phi^*(a_i)]\right) = \mathbb{P}[\phi^* = 1] + \frac{2\alpha}{d^*} \sum_{i=1}^{d^*/2} \mathbf{1}[\phi(a_i) \neq \phi^*(a_i)].$$

对于任意样本 $S \in S^m$，令 $N(S) = (N_1(S), ..., N_{d^*/2}(S))$，其中 $N_i(S)$ 是 S 中 a_i 或 b_i 出现的次数。于是，令 L 为 Φ' 的学习算法，我们有，对于 $\phi := L(S)$，

$$\begin{aligned}\mathbb{E}\left[\frac{2}{d^*} \sum_{i=1}^{d^*/2} \mathbf{1}[\hat{\phi}(a_i) \neq \phi^*(a_i)]\right] &= \frac{2}{d^*} \mathbb{E}\left[\sum_{i=1}^{d^*/2} \mathbf{1}[\hat{\phi}(a_i) \neq \phi^*(a_i)]\right] \\ &= \frac{2}{d^*} \sum_{N} \sum_{i=1}^{d^*/2} \mathbb{P}[\hat{\phi}(a_i) \neq \phi^*(a_i) \mid N(S) = N] \cdot \mathbb{P}[N(S) = N].\end{aligned}$$

其中 $N = (N_1, ..., N_{d^*/2})$ 遍历满足 $\sum_{i=1}^{d^*/2} N_i = m$ 的所有 $d^*/2$ 元正整数元组集合。

类似于 Anthony 和 Bartlett（2009年，定理 5.2）的证明，我们可以得出结论：若 $m < \frac{d^*}{320 \cdot \varepsilon^2}$，则对于样本 $S \sim \mathcal{D}^m$，以大于 $1/64$ 的概率有 $\mathbb{P}[\phi(X) = 1] - \mathbb{P}[\phi^*(X) = 1] > \varepsilon$.。特别地，这对 $\hat{\phi} = \operatorname{argmin}_{\phi \in \Phi'} \mathbb{P}[\phi(X) = 1]$ 成立。

设 $m < \frac{d^*}{1280 \cdot \varepsilon^2}$。根据上述结论，对于样本 $S \sim \mathcal{D}^m$，以大于 δ 的概率有 $\mathbb{P}[\hat{\phi}(X) = 1] - \mathbb{P}[\phi^*(X) = 1] > 2\varepsilon$.

我们现在声称，以下至少一项成立的概率大于 δ：

- $|\mathbb{P}[\hat{\phi}(X) = 1] - \mathbb{P}[\phi(X) = 1]| > \varepsilon$; ˆ
- $|\mathbb{P}[\phi^*(X) = 1] - \mathbb{P}[\phi^*(X) = 1]| > \varepsilon$。

反设这些不等式中的每一个成立的概率至多为 δ。那么，以至少$1 - 2\delta$的概率,

$$\mathbb{P}[\hat{\phi}(X) = 1] - \varepsilon < \hat{\mathbb{P}}[\hat{\phi}(X) = 1$$

and

$$\hat{\mathbb{P}}[\phi^*(X) = 1] < \mathbb{P}[\phi^*(X) = 1] + \varepsilon.$$

此外，根据 ˆ ϕ 的定义，我们有 ˆ $\mathbb{P}[\hat{\phi}(X) = 1] \leq \hat{\mathbb{P}}[\phi^*(X) = 1]$. 我们得到，以至少 $1 - 2\delta$ 的概率，$\mathbb{P}[\hat{\phi}(X) = 1] - \mathbb{P}[\phi^*(X) = 1] < 2\varepsilon$。由于 $\delta < 1/128$ 和 $m < \frac{d^*}{1280 \cdot \varepsilon^2}$，这与上述下界相矛盾。因此，前述假设的不等式中至少有一个成立，从而证明了该命题。 □

A.2 充分条件的证明

We 现在证明定理 5.5。我们推导出使下列不等式成立的一个充分条件:

$$\mathrm{B}_{\mathrm{PR}} = \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \hat{\varepsilon}_{\mathrm{ERM}} + R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}}, d) + R_{\mathrm{f}}(d) = \mathrm{B}_{\mathrm{ERM}}.$$

在假设 ˆ$\varepsilon_{\mathrm{ERM}} = \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u$ 的条件下，只需满足

$$R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}}, d) + R_{\mathrm{f}}(d).$$

这等价于

$$\sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + \sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) + R_{\mathrm{f}}(d).$$

由于 $\widehat{\varepsilon}_{ig} \leq$ ˆ$\varepsilon_{\mathrm{ERM}}$，只需满足

$$\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d^*) + \sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) + R_{\mathrm{f}}(d),$$

这等价于

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{R_{\mathrm{s}}(1, d) - R_{\mathrm{s}}(1, d^*)}{R_{\mathrm{s}}(1, d_a)} + \frac{R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a)}{R_{\mathrm{s}}(1, d_a)}.$$

对于固定的 δ，当 d、$d*$、d_a 足够大时，这等价于

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \Theta(\frac{\sqrt{d} - \sqrt{d^*}}{\sqrt{d_a}}) - \sqrt{\frac{\log(m)}{m}} \cdot \Theta(\frac{d_a + d^* - d}{\sqrt{d_a}}),$$

如所述。

A.3 必要条件的证明

现在我们证明定理 5.6。首先，我们证明一个额外的引理，它给出了 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$ 的必要条件。引理 A.1。对于任意固定的 $\delta \in \in$(0,1)，若 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$，则

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{\sqrt{d}}{\sqrt{d_a}} - \sqrt{\widehat{\varepsilon}_{ig}} \cdot \frac{\sqrt{d^*}}{\sqrt{d_a}} + o(1),$$

最后一项的收敛是相对于 d、d^* 一同增长和/或 m 增长而言的。

证明。根据该引理的假设，$\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$。因此，根据定义,

$$\widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \hat{\varepsilon}_{\mathrm{ERM}} + \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) + R_{\mathrm{f}}(d).$$

这等价于:

$$\widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + \sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \hat{\varepsilon}_{\mathrm{ERM}} + \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) + R_{\mathrm{f}}(d).$$

因此,

$$\begin{aligned}\sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) &\leq \hat{\varepsilon}_{\mathrm{ERM}} - (\widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u) + \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d_a) - R_{\mathrm{f}}(d^*) \\ &\leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a).\end{aligned}$$

最后一个不等式由引理 5.4 得出，$\hat{\varepsilon}_{\mathrm{ERM}} \leq \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u$。

现在，根据 R_{f} 的定义，我们有

$$R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a) = 8\frac{\log(m+1)}{m}(d - d^* - d_a) - \frac{4\log(\frac{4}{\delta})}{m}.$$

根据引理 5.1，$d_a \geq d + d^* - 2$。因此，$d - d^* - d_a < 0$。由此可得 $R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a) < 0$。因此，

$$\sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*).$$

由此可得

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*)}{R_{\mathrm{s}}(1, d_a)} = \frac{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}(d + A)} - \sqrt{\widehat{\varepsilon}_{ig}(d^* + A)}}{\sqrt{d_a + A}},$$

其中 $A := \log(4/\delta)/(2\log(m+1))$。因此,

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{\sqrt{d}}{\sqrt{d_a}} - \sqrt{\widehat{\varepsilon}_{ig}} \cdot \frac{\sqrt{d^*}}{\sqrt{d_a}} + o(1),$$

其中最后一项的收敛是相对于 d, d^* 同时增长和/或 m 增长而言的。 □

接下来，我们利用上述两个引理证明该定理。

定理 5.6 的证明。假设 $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$。根据引理 A.1,

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{\sqrt{d}}{\sqrt{d_a}} - \sqrt{\widehat{\varepsilon}_{ig}} \cdot \frac{\sqrt{d^*}}{\sqrt{d_a}} + o(1).$$

记 $\alpha := d^*/d_a$。假设 $\alpha \geq 1$（否则定理中的陈述显然成立）。我们有

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{\sqrt{d}}{\sqrt{d_a}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1).$$

这里的收敛是 在固定的 α 下，随着 d、d^* 或 m 的增长而发生。由于 $d_a \geq d^* + d - 2 = (1+\alpha) \cdot d - 2$，我们有

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{1}{\sqrt{1+\alpha}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1). \tag{11}$$

由于 $\sqrt{\widehat{\varepsilon}_u} \geq 0$，我们有

$$\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}} + o(1) \geq 0.$$

因此，$\widehat{\varepsilon}_{ig} \leq \hat{\varepsilon}_{\mathrm{ERM}}/\alpha + o(1)$。结合引理 5.4，我们得到

$$\hat{\varepsilon}_{\mathrm{ERM}} \leq \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u \leq \hat{\varepsilon}_{\mathrm{ERM}}/\alpha + \widehat{\varepsilon}_u + o(1).$$

结合公式（11），可得

$$\sqrt{(1 - \frac{1}{\alpha}) \cdot \hat{\varepsilon}_{\mathrm{ERM}}} \leq \sqrt{\widehat{\varepsilon}_u} \leq \frac{1}{\sqrt{1+\alpha}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1).$$

重新整理后，得到

$$\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})\leq 1-\sqrt{\alpha}\cdot\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}+o(1).$$

由此得到

$$\sqrt{\alpha}\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}\leq 1-\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})+o(1).$$

由于 $0\leq\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}$，必有 $\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})\leq 1+o(1)$。解出 α，得到 $\alpha\leq 2.25+o(1)$。由于 $\alpha=d^*/d$，我们得出 $d^*\leq 2.25\cdot d+o(1)$，如所声称。

□