1 The Probit Model

Suppose we are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ containing n pairs of observations. We assume that the x_i are d-dimensional vectors, i.e. $x_i \in \mathbb{R}^d$ that contain explanatory information regarding the binary outcome $y_i \in \{0, 1\}$. Perhaps the x_i are used to represent information about a patient, such as blood pressure or weight, and the y_i are used to indicate the presence or absence of a heart disease. In such a setting we are often interested in modeling the relationship between explanatory values of x_i and the binary outcomes y_i.

For convenience, we put all the observations x_i inside of a matrix $X \in \mathbb{R}^{n\times d}$ in such a way that the i-th row of X corresponds to x_i. We do the same with the values of y_i and put them in a vector $y \in \{0, 1\}^n$.

Since it is reasonable to assume that there is a degree of randomness involved in the data generating process, we model the y_i as realizations of independent random variables Y_i that can be summarized as a single random vector Y, where Y_i is the i-th component of Y. We will use this upper case notation in the following to distinguish between random variables and their realizations. This brings us to the first assumption in the probit model: We assume that the observations are independent, i.e. their outcomes don't influence each other.

The second assumption of the probit model is that there is a hidden random quantity Y_i^* that is associated with each outcome Y_i in that it directly determines its result like this:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{if } Y_i^* \leq 0 \end{cases} \tag{1}$$

These Y_i^*, that can also be summarized as a random vector Y^*, are also assumed to be independent and, as already noted, unobservable. The third assumption of the probit model is, that the observed values x_i influence Y_i^* in the form of a classical linear model:

$$Y^* = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

where $\beta \in \mathbb{R}^d$ is the parameter vector of the linear model, ϵ is a normal distributed vector with independent components of mean zero and variance σ^2, and $I \in \mathbb{R}^{n\times n}$ is the $n \times n$ identity matrix. It follows directly that Y^* is also normal distributed: $Y^* \sim \mathcal{N}(X\beta, \sigma^2 I)$.

These three assumptions are already a complete specification of the probit model and are summarized in the following definition as a brief recapitulation:

Definition 1 (Probit Model)**.** *A dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with model matrix $X \in \mathbb{R}^{n\times d}$ and observed response vector $y \in \{0, 1\}^n$ was generated by a probit model with parameters $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_{>0}$, if the following three assumptions are true:*

1. *The observations $y_1, ..., y_n$ are realizations of independent binary random variables $Y_1, ..., Y_n$.*

2. *The outcomes of $Y_1, ..., Y_n$ are determined by hidden continuous random variables $Y_1^*, ..., Y_n^*$ by thresholding: If $Y_i^* > 0$, then $Y_i = 1$, and if $Y_i^* \leq 0$, then $Y_i = 0$.*

3. *The vector of hidden variables Y^* follows a multivariate normal distribution: $Y^* \sim \mathcal{N}(X\beta, \sigma^2 I)$, where $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_{>0}$ are the model parameters.*

From this definition, it is straight forward to determine the distribution of the response variables Y_i. We can calculate the probability $P(Y_i = 1)$ like this:

$$P(Y_i = 1) = P(Y_i^* > 0) = 1 - P(Y_i^* \leq 0) = 1 - P\left(\frac{Y_i^* - x_i\beta}{\sigma} \leq -\frac{x_i\beta}{\sigma}\right) = \Phi\left(\frac{x_i\beta}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

The result $P(Y_i = 1) = \Phi\left(\frac{x_i\beta}{\sigma}\right)$ leads us to an interesting observation: Both parameters β and σ are unknown model parameters and every value of σ can be compensated by a corresponding scaling of β. This means that, because we can't observe the hidden variables Y_i^*, it is impossible to determine which β and which σ generated the data without any prior knowledge. We can only draw conclusions with regard to the scaled parameter $\frac{1}{\sigma}\beta$. In this situation, we say that β and σ are *not identifiable.*

For this reason, we can assume without losing generality, that $\sigma = 1$ and arrive at

$$P(Y_i = 1) = \Phi(x_i\beta). \tag{3}$$

Since Y_i is binary, it follows that

$$P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \Phi(x_i\beta) = \Phi(-x_i\beta),$$

which immediately leads us to the model equations

$$Y_i \sim Bin(1, \pi_i), \quad \pi_i = \Phi(x_i\beta). \tag{4}$$

1.1 The probit model is a special case of the generalized linear model

2 Coresets

Definition 2. *Let $X \in \mathbb{R}^{n \times d}$, $y \in \{-1,1\}^n$ be an instance of probit regression with sample weights $w \in \mathbb{R}^n_{>0}$ and let $z_i = -y_i x_i^T$, $i = 1, ..., n$. Then $C \in \mathbb{R}^{k \times d}$ weighted by $u \in \mathbb{R}^k_{>0}$ is a $(1 \pm \epsilon)$-coreset of X, y for probit regression if*

$$(1-\epsilon) f_{w,Z}(\beta) \leq f_{u,C}(\beta) \leq (1+\epsilon) f_{w,Z}(\beta) \quad \forall \beta \in \mathbb{R}^d,$$

where $f_{w,Z}(\beta) = \sum_{i=1}^n w_i g(z_i \beta)$, $f_{u,C}(\beta) = \sum_{i=1}^k u_i g(c_i \beta)$ and $g(z) = -\log \Phi(-z)$.

2.1 Lower Bounds

Theorem 1. *Let $X \in \mathbb{R}^{n \times 2}$, $y \in \{-1,1\}^n$ be an instance of probit regression. Any coreset $C \in \mathbb{R}^{k \times 2}$ of X, y for probit regression consists of at least $k \in \Omega\left(\frac{n}{\log n}\right)$ points.*

Proof. We first show how such a coreset could be used in a communication protocol for the INDEX communication game to encode a message. Since there exists a lower bound on the minimum message length of the INDEX game (see [Kremer et al., 1999]), we can use it to derive a lower bound on the coreset size. The same technique was also used in [Munteanu et al., 2018] to find lower bounds for coresets of logistic regression and is here slightly adapted for probit regression.

The INDEX game consists of two players, Alice and Bob. Alice is given a random binary string $x \in \{0,1\}^n$ of n bits and Bob is given an index $i \in [n]$. The goal is for Alice to send a message to Bob that allows Bob to obtain the value x_i of Alice's binary string x. It was shown in [Kremer et al., 1999], that the minimum length of a message sent by Alice that still allows Bob to obtain x_i with constant probability is in $\Omega(n)$ bits. We will now see how a coreset for probit regression can be used to encode such a message.

The first step is for Alice to convert her binary string x into a set P of two-dimensional points as follows: For each entry x_j of her binary string where $x_j = 1$, she adds a point $p_j = \left(\cos\left(2\pi \frac{j}{n}\right), \sin\left(2\pi \frac{j}{n}\right)\right)$ to her set P and labels it with 1. As we can see, all of these points are on the unit circle and all of them are labeled with 1. Next, she uses these points to construct a coreset for probit regression $C \in \mathbb{R}^{k \times 2}$ of P and sends it to Bob. We will later see, how large the size k of this coreset must be, so that Bob can still obtain x_i with constant probability.

As soon as Alice's coreset C arrives at Bob, Bob can use it to obtain the value of x_i. To do this, Bob first adds two new points $q_1 = \left(\cos\left(2\pi \frac{i-0.5}{n}\right), \sin\left(2\pi \frac{i-0.5}{n}\right)\right)$ and $q_2 = \left(\cos\left(2\pi \frac{i+0.5}{n}\right), \sin\left(2\pi \frac{i+0.5}{n}\right)\right)$ to the set and labels both points with -1 (see figure 1). Next, he uses his points q_1 and q_2 together with the coreset C to obtain a solution for the corresponding probit regression problem. He can then use the value of the cost function to determine the value of x_i like this:

Since Alice only added a point p_j to her set if $x_j = 1$, both of his points q_1 and q_2 are linearly seperable from Alice's points if the value of $x_i = 0$, i.e. Alice didn't add a point for x_i. In this case, the value of the cost function tends to zero. If, on the other hand, Bob's new points q_1 and q_2 can't be linearly seperated from the other

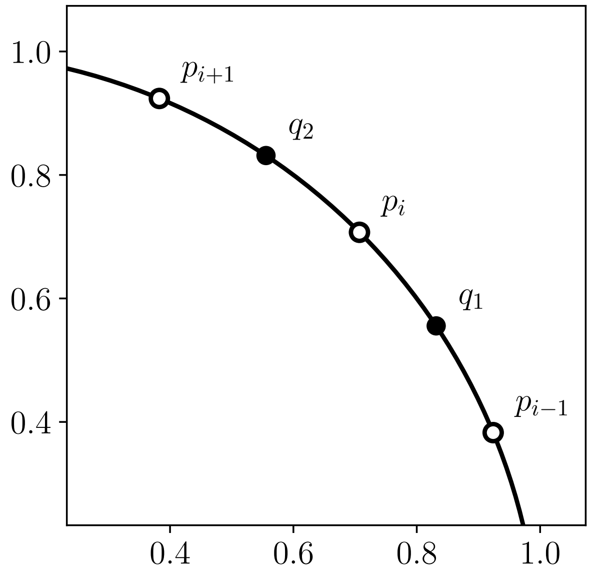

Figure 1: Bob places two points q_1 and q_2 in such a way on the unit circle, that they can be linearly seperated from the other points if and only if Alice didn't place a point at p_i.

points, it means that Alice added a point for $x_i = 1$. In this case, there must be at least one misclassification and the value of the cost function is at least $g(0) = \log(2)$. Since coresets can be used to obtain $(1 + \epsilon)$-approximation of the objective function, Bob can use this case distinction to determine the value of x_i.

There is one special case that has to be dealt with in order for this protocol to work. If Alice's coreset only consists of the single point p_i, Bob's points q_1 and q_2 could still be linearly seperated although Alice added p_i. The workaround to this is simple though: Bob can always just add two more points at the locations of p_{i-1} and p_{i+1} and label them with 1. Now, q_1 and q_2 can only be linearly seperated from the other points if and only if Alice didn't add a point p_i.

Let us now see how large the size k of Alice's coreset must be for this protocol to work with constant probability. In [Kremer et al., 1999] it was shown, that the minimum length of a message that Alice must send is in $\Omega(n)$ bits. Since each of the points that Alice created can be encoded in $\log(n)$ space, it follows from the lower bound that $\Omega(n) \subseteq \Omega(k \log(n))$, so k must be in $\Omega\left(\frac{n}{\log(n)}\right)$.

We can conclude that if there existed a $(1 + \epsilon)$-coreset for probit regression with size $k \in o\left(\frac{n}{\log(n)}\right)$, it would contradict the minimum message length of INDEX, which proves the claim. □

3 Sensitivity Sampling

Definition 3. *Let $Z \in \mathbb{R}^{n \times d}$. Then we define*

$$\mu_w(Z) = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\left\|(\sqrt{D_w}Z\beta)^+\right\|_2^2}{\left\|(\sqrt{D_w}Z\beta)^-\right\|_2^2} = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\left\|(\sqrt{D_w}Z\beta)^-\right\|_2^2}{\left\|(\sqrt{D_w}Z\beta)^+\right\|_2^2}$$

Z weighted by w is called μ-complex if $\mu_w(Z) \leq \mu$.

Definition 4 ([Feldman et al., 2020, Langberg and Schulman, 2010])**.** *Let $F = \{g_1, ..., g_n\}$ be a set of functions, $g_i : \mathbb{R} \to \mathbb{R}^{\geq 0}$, $i = 1, ..., n$ weighted by $w \in \mathbb{R}^n_{>0}$. The sensitivity of g_i for $f_w(\beta) = \sum_{i=1}^n w_i g_i(\beta)$ is defined as*

$$\varsigma_i = \sup_{\beta \in \mathbb{R}^d,\ f_w(\beta) > 0} \frac{w_i g_i(\beta)}{f_w(\beta)}.$$

The total sensitivity, i.e. the sum of the sensitivities is $\mathfrak{S} = \sum_{i=1}^n \varsigma_i$.

Definition 5 ([Feldman et al., 2020])**.** *A range space is a pair $\mathfrak{R} = (F, \mathcal{R})$, where F is a set and $\mathcal{R}$ is a family (set) of subsets of F, called ranges.*

Definition 6 ([Feldman et al., 2020])**.** *The VC-dimension $\Delta(\mathfrak{R})$ of a range space $\mathfrak{R} = (F, \mathcal{R})$ is the size $|G|$ of the largest subset $G \subseteq F$ such that*

$$|\{G \cap \text{range} \mid \text{range} \in \mathcal{R}\}| = 2^{|G|},$$

i.e. G is shattered by $\mathcal{R}$.

Definition 7 ([Feldman et al., 2020])**.** *Let F be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}^{\geq 0}$. For every $\beta \in \mathbb{R}^d$ and $r \geq 0$, let*

$$\text{range}(F, \beta, r) = \{f \in F \mid f(\beta) \geq r\}$$

and let

$$\mathcal{R}(F) = \left\{\text{range}(F, \beta, r) \mid \beta \in \mathbb{R}^d,\ r \geq 0\right\}.$$

Then we call $\mathfrak{R}_F := (F, \mathcal{R}(F))$ the range space induced by F.

Theorem 2 ([Feldman et al., 2020, Munteanu et al., 2018])**.** *Let $\mathcal{F} = \{f_1, ..., f_n\}$ be a set of functions, $f_i : \mathbb{R} \to \mathbb{R}^{\geq 0}$, $i = 1, ..., n$ weighted by $w \in \mathbb{R}^n_{>0}$. Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $s_i \geq \varsigma_i$. Let $S = \sum_{i=1}^n s_i \geq \mathfrak{S}$. Given s_i, one can compute in time $O(|\mathcal{F}|)$ a set $\mathcal{R} \subseteq \mathcal{F}$ of*

$$O\left(\frac{S}{\epsilon^2}\left(\Delta \log S + \log\left(\frac{1}{\delta}\right)\right)\right)$$

weighted functions such that with probability $1 - \delta$ we have for all $\beta \in \mathbb{R}^d$ simultaneously

$$\left|\sum_{f \in \mathcal{F}} w_i f_i(\beta) - \sum_{f \in \mathcal{R}} u_i f_i(\beta)\right| \leq \epsilon \sum_{f \in \mathcal{F}} w_i f_i(\beta)$$

where each element of $\mathcal{R}$ is sampled independently with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$, $u_i = \frac{Sw_j}{s_j|\mathcal{R}|}$ denotes the weight of a function $f_i \in \mathcal{R}$ that corresponds to $f_j \in \mathcal{F}$, and where Δ is an upper bound on the VC-dimension of the range space $\mathfrak{R}_{\mathcal{F}^*}$ induced by $\mathcal{F}^*$. $\mathcal{F}^*$ is the set of functions $f_j \in \mathcal{F}$ scaled by $\frac{Sw_j}{s_j|\mathcal{R}|}$.

Lemma 1. *Let $Z \in \mathbb{R}^{n \times d}$, $c \in \mathbb{R}_{>0}$. The range space induced by*

$$\mathcal{F}^c_{probit} = \{cg(z_i\beta) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}^c_{probit}}) \leq d+1$.

Proof. For all $G \subseteq \mathcal{F}^c_{probit}$ we have

$$\left|\left\{G \cap \text{range} \mid \text{range} \in \mathcal{R}(\mathcal{F}^c_{probit})\right\}\right| = \left|\left\{\text{range}(G,\beta,r) \mid \beta \in \mathbb{R}^d,\ r \geq 0\right\}\right|.$$

Since g is invertible and monotone, we have for all $\beta \in \mathbb{R}^d$ and $r \geq 0$ that

$$\begin{aligned}\text{range}(G,\beta,r) &= \{g_i \in G \mid g_i(\beta) \geq r\} \\ &= \{g_i \in G \mid cg(x_i\beta) \geq r\} \\ &= \left\{g_i \in G \mid x_i\beta \geq g^{-1}\left(\frac{r}{c}\right)\right\}.\end{aligned}$$

Note, that $\left\{g_i \in G \mid x_i\beta \geq g^{-1}\left(\frac{r}{c}\right)\right\}$ corresponds to the positively classified points of the affine hyperplane classifier $x \mapsto \text{sign}\left(x\beta - g^{-1}\left(\frac{r}{c}\right)\right)$. We thus have for all $G \subseteq \mathcal{F}^c_{probit}$, that

$$\left|\left\{G \cap \text{range} \mid \text{range} \in \mathcal{R}(\mathcal{F}^c_{probit})\right\}\right| = \left|\left\{\{g_i \in G \mid x_i\beta - s \geq 0\} \ \mid \beta \in \mathbb{R}^d,\ s \in \mathbb{R}\right\}\right|.$$

Since the VC dimension of the set of affine hyperplane classifiers is $d+1$, it follows that $\Delta(\mathfrak{R}_{\mathcal{F}^c_{probit}}) \leq d+1$, which concludes our proof. □

Lemma 2. *Let $Z \in \mathbb{R}^{n \times d}$ be weighted by $w \in \mathbb{R}^n_{>0}$ where $w_i \in \{v_1, ..., v_t\}$ for all $i \in [n]$. The range space induced by*

$$\mathcal{F}_{probit} = \{w_i g(z_i\beta) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}_{probit}}) \leq t \cdot (d+1)$.

Proof. We partition the functions of $\mathcal{F}_{probit}$ into t disjoint classes

$$F_j = \{w_i g(z_i\beta) \in \mathcal{F}_{probit} \mid w_i = v_j\}, \quad j \in [t].$$

The functions in each of these classes have an equal weight, wich means that by lemma 1, each of their induced range spaces has a VC-dimension of at most $d+1$.

For the sake of contradiction, assume that $\Delta(\mathfrak{R}_{\mathcal{F}_{probit}}) > t \cdot (d+1)$ and let G be the corresponding set of size $|G| > t \cdot (d+1)$ that is shattered by $\mathcal{R}(\mathcal{F}_{probit})$. Since the sets F_j are disjoint, each intersection $F_j \cap G$ must be shattered by $\mathcal{R}(F_j)$. Further, at least one of the intersections must have at minimum $\frac{|G|}{t}$ elements, which means that for at least one $j \in [t]$ it holds that $|F_j \cap G| \geq \frac{|G|}{t} > \frac{t \cdot (d+1)}{t} = d+1$. This is a contradiction to lemma 1, which concludes the proof. □

Lemma 3. *Let $Z \in \mathbb{R}^{n \times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. If for index i, the supreme β in definition 4 satisfies $2 \leq z_i\beta$, then $w_i g(z_i\beta) \leq 2\|U_i\|_2^2(1+\mu)f_w(\beta)$.*

Proof. Let $\sqrt{D_w}Z = UR$, where U is an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. It follows from $2 \leq z_i\beta$ and from the monotonicity of g that

$$\begin{aligned}
w_i g(z_i\beta) &= w_i g\left(\frac{\sqrt{w_i}z_i\beta}{\sqrt{w_i}}\right) = w_i g\left(\frac{U_iR\beta}{\sqrt{w_i}}\right) \leq w_i g\left(\frac{\|U_i\|_2\|R\beta\|_2}{\sqrt{w_i}}\right) \\
&= w_i g\left(\frac{\|U_i\|_2\|UR\beta\|_2}{\sqrt{w_i}}\right) = w_i g\left(\frac{\|U_i\|_2\|\sqrt{D_w}Z\beta\|_2}{\sqrt{w_i}}\right) \\
&\leq \|U_i\|_2^2\|\sqrt{D_w}Z\beta\|_2^2 \leq \|U_i\|_2^2(1+\mu)\|(\sqrt{D_w}Z\beta)^+\|_2^2 \\
&= \|U_i\|_2^2(1+\mu)\sum_{j:\ \sqrt{w_j}z_j\beta \geq 0} w_j(z_j\beta)^2 \\
&\leq 2\|U_i\|_2^2(1+\mu)\sum_{j:\ \sqrt{w_j}z_j\beta \geq 0} w_j g(z_j\beta) \\
&\leq 2\|U_i\|_2^2(1+\mu)\sum_{j=1}^{n} w_j g(z_j\beta) \\
&= 2\|U_i\|_2^2(1+\mu)f_w(\beta)
\end{aligned}$$

□

Lemma 4. *Let $Z \in \mathbb{R}^{n \times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. If for index i, the supreme β in definition 4 satisfies $z_i\beta \leq 2$, then $w_i g(z_i\beta) \leq \frac{w_i}{\mathcal{W}}(80+16\mu)f_w(\beta)$.*

Proof. Let $K^- = \{j \in [n] \mid z_j\beta \leq -1\}$ and $K^+ = \{j \in [n] \mid z_j\beta > -1\}$. Note that $g(-1) > \frac{1}{10}$ and $g(z_i\beta) \leq g(2) < 4$. Also, $\sum_{j \in K^+} w_j + \sum_{j \in K^-} w_j = \mathcal{W}$.
Thus, if $\sum_{j \in K^+} w_j \geq \frac{1}{2}\mathcal{W}$ then

$$f_w(\beta) = \sum_{j=1}^{n} w_j g(z_j\beta) \geq \sum_{j \in K^+} w_j g(z_j\beta) \geq \frac{\sum_{j \in K^+} w_j}{10} \geq \frac{\mathcal{W}}{20} = \frac{\mathcal{W}}{20w_i}w_i \geq \frac{\mathcal{W}}{80w_i}w_i g(z_i\beta)$$

If on the other hand $\sum_{j\in K^+} w_j < \frac{1}{2}\mathcal{W}$, then $\sum_{j\in K^-} w_j \geq \frac{1}{2}\mathcal{W}$. Thus

$$\begin{aligned} f_w(\beta) &= \sum_{j=1}^{n} w_j g(z_j\beta) \geq \sum_{j:\ z_j\beta>0} w_j g(z_j\beta) \geq \frac{1}{2} \sum_{j:\ z_j\beta>0} w_j (z_j\beta)^2 \\ &= \frac{1}{2}\|(\sqrt{D_w}Z\beta)^+\|_2^2 \geq \frac{1}{2\mu}\|(\sqrt{D_w}Z\beta)^-\|_2^2 \\ &= \frac{1}{2\mu} \sum_{j:\ z_j\beta<0} w_j (z_j\beta)^2 \\ &\geq \frac{1}{2\mu} \sum_{j\in K^-} w_j (z_j\beta)^2 \\ &\geq \frac{1}{2\mu} \sum_{j\in K^-} w_j \\ &\geq \frac{\mathcal{W}}{4\mu} \\ &\geq \frac{\mathcal{W}}{16\mu w_i} w_i g(z_i\beta) \end{aligned}$$

Adding both bounds, we get that for $z_i\beta \leq 2$:

$$w_i g(z_i\beta) \leq f_w(\beta)\frac{80 w_i}{\mathcal{W}} + f_w(\beta)\frac{16\mu w_i}{\mathcal{W}} = \frac{w_i}{\mathcal{W}}(80+16\mu) f_w(\beta)$$

□

Lemma 5. *Let $Z \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. For each $i \in [n]$, the sensitivity of $g_i(\beta) = g(z_i\beta)$ is bounded by $\varsigma_i \leq s_i = (80+16\mu)(\|U_i\|_2^2 + \frac{w_i}{\mathcal{W}})$. The total sensitivity is bounded by $\mathfrak{S} \leq 192\mu d$.*

Proof.

$$\begin{aligned} \varsigma_i &= \sup_\beta \frac{w_i g(z_i\beta)}{f_w(\beta)} \leq \sup_\beta \frac{2\|U_i\|_2^2(1+\mu) f_w(\beta) + \frac{w_i}{\mathcal{W}}(80+16\mu) f_w(\beta)}{f_w(\beta)} \\ &= 2\|U_i\|_2^2(1+\mu) + \frac{w_i}{\mathcal{W}}(80+16\mu) \\ &\leq \|U_i\|_2^2(80+16\mu) + \frac{w_i}{\mathcal{W}}(80+16\mu) \\ &= (80+16\mu)(\|U_i\|_2^2 + \frac{w_i}{\mathcal{W}}) \end{aligned}$$

$$
\begin{aligned}
\mathfrak{S} = \sum_{i=1}^{n} \varsigma_i &\leq (80+16\mu) \sum_{i=1}^{n} \|U_i\|_2^2 + \frac{w_i}{\mathcal{W}} \\
&= (80+16\mu)(\|U\|_F^2 + 1) \\
&= (80+16\mu)(d+1) \\
&\leq 96\mu(d+1) \\
&\leq 192\mu d
\end{aligned}
$$

□

Lemma 6. *Let $U \in \mathbb{R}^{n \times d}$ be an orthonormal matrix. Then $\|U\|_F^2 = d$.*

Proof.

$$
\begin{aligned}
\|U\|_F^2 &= \sum_{i=1}^{n} \sum_{j=1}^{d} |u_{ij}|^2 \\
&= \sum_{j=1}^{d} \sum_{i=1}^{n} |u_{ij}|^2 \\
&\overset{(1)}{=} \sum_{j=1}^{d} 1 \\
&= d
\end{aligned}
$$

(1) follows from the fact that the columns of U have unit norm due to its orthonormality.

□

4 Notes

4.1 VC Dimension

An alternative approach is to write down the VC dimension by using an instance space and a concept class as given in [Kearns and Vazirani, 1994].

Lemma 7. *Let $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d \times \mathbb{R}_{>0}$ be the instance space consisting of n points with their last coordinate being positive. The concept class of interest, $\mathcal{C}$ over X, is given as follows:*

$$\mathcal{C} = \left\{ \{x \in X : \ f_{\beta,r}(x) \geq 0\} \mid \beta \in \mathbb{R}^d, r \geq 0 \right\},$$

with

$$f_{\beta,r}(x) = x_{d+1} \cdot g\left(\sum_{i=1}^{d} x_i \beta_i\right) - r$$

and

$$g(x) = -\log \Phi(-x).$$

The VC dimension of $\mathcal{C}$ is equal to the VC dimension of the range space induced by $\mathcal{F}^w_{probit} = \{w_i g(z_i \beta) \mid i \in [n]\}$, $Z \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^n_{>0}$.

There a few different strategies that can be used to find an upper bound on the VC dimension of $\mathcal{C}$, as shown by the following lemmas. The first one is a simple upper bound for finite concept classes:

Lemma 8. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. If the cardinality of $\mathcal{C}$ can be bounded by m, i.e. $|\mathcal{C}| \leq m$, then $VCdim(\mathcal{C}) \leq log(m)$.*

The next lemma partitions the concept class into smaller classes, for each of which the VC dimension can be bounded:

Lemma 9. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{C}$ into k disjoint subsets, i.e. $\mathcal{C} = \bigcup_{i=1}^k \mathcal{C}_i$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset \ \forall i \neq j$. Then, $VCdim(\mathcal{C}) \leq \sum_{i=1}^k VCdim(\mathcal{C}_i)$.*

Proof. For the sake of contradiction, assume there was a set $S \subseteq X$ of size $|S| > \sum_{i=1}^k VCdim(\mathcal{C}_i)$ that is shattered by $\mathcal{C}$. If S is shattered by C, every subset of S must also be shattered by C. Consider the intersections $T_i = \bigcup_{c \in \mathcal{C}_i} S \cap c$. Every T_i is a subset of S and $S = \bigcup_{i=1}^k T_i$. Since S is shattered by $\mathcal{C}$, every T_i must be shattered by $\mathcal{C}_i$. We assumed that $|S| > \sum_{i=1}^k VCdim(\mathcal{C}_i)$. It follows that there exists a T_j with $T_j > VCdim(\mathcal{C}_j)$. Since T_j is also shattered by $\mathcal{C}_j$, this is a contradiction, which concludes the proof. □

A result in [Linial et al., 1991] suggests an even smaller upper bound:

Lemma 10 ([Linial et al., 1991])**.** *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $\mathcal{C} = \bigcup_{i=1}^k \mathcal{C}_i$ and $VCdim(\mathcal{C}_i) \leq m$. If k is bounded by a polynomial function of m, then $VCdim(\mathcal{C}) \leq 3m$.*

Instead of partitioning the concept class, we could also partition the instance space and obtain a similar bound:

Lemma 11. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $X_1, ..., X_k$ be a partition of X into k disjoint subsets, i.e. $X = \bigcup_{i=1}^{k} X_i$ and $X_i \cap X_j = \emptyset \ \forall i \neq j$. Let $\mathcal{C}_i = \{X_i \cap c \mid c \in \mathcal{C}\}$ be a concept class over X_i for all $i \in [k]$.*
Then, $VCdim(\mathcal{C}) \leq \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$.

Proof. Again, assume there existed a set $S \subseteq X$ of size $|S| > \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$ that is shattered by $\mathcal{C}$. S can be partitioned into disjoined subsets $T_i = S \cap X_i$, with $\bigcup_{i=1}^{k} T_i = S$. Every T_i must be shattered by $\mathcal{C}_i$. Since we assumed that $|S| > \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$, there exists a T_j with $|T_j| > VCdim(C_j)$ which is also shattered by C_j. This contradiction concludes the proof. □

4.2 New idea for VC dimension proof

Lemma 12. *Let*

$$h_{\beta,r}(x) = \begin{cases} 1 \text{ if} \quad x_{d+1} \cdot g\left(\sum_{i=1}^{d} x_i\beta_i\right) - r \geq 0 \\ 0 \text{ else} \end{cases}$$

Be a function from $\mathbb{R}^{d+1}$ *to* $\{0, 1\}$ *with parameters* $\beta \in \mathbb{R}^d$ *and* $r \in \mathbb{R}_{\geq 0}$ *with*

$$g(x) = \log\left(\frac{1}{1 - \Phi(x)}\right),$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} dz \ .$$

Let

$$H = \left\{x \mapsto h_{\beta,r}(x) \mid \beta \in \mathbb{R}^d, \ r \in \mathbb{R}_{\geq 0}\right\}$$

be the hypothesis class determined by h*. Then, the VC dimension of* H *is ...*

Proof. Let $S = \sum_{i=1}^{d} x_i\beta_i$. We show that h can be computed in t steps as follows:

$$\begin{aligned} & x_{d+1} \cdot g\left(S\right) - r \geq 0 \\ \iff & \log\left(\frac{1}{1 - \Phi(S)}\right) \geq \frac{r}{x_{d+1}} \\ \iff & \frac{1}{1 - \Phi(S)} \geq \exp\left(\frac{r}{x_{d+1}}\right) \\ \iff & 1 - \Phi(S) \leq \exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \Phi(S) \geq 1 - \exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{S} e^{-\frac{1}{2}z^2} dz \geq 1 - \exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \int_{-\infty}^{S} e^{-\frac{1}{2}z^2} dz \geq \sqrt{2\pi}\left(1 - \exp\left(-\frac{r}{x_{d+1}}\right)\right) \end{aligned}$$

□

4.3 Online Leverage Scores

The leverage scores of a matrix $A \in \mathbb{R}^{n \times d}$ are given by $l_i = a_i^T (A^T A)^{-1} a_i$ [Cohen et al., 2020]. According to [Cohen et al., 2020], we can obtain overestimates of these scores by using only a subset of the rows in A to compute them.

Let A_j be a matrix that contains only the first j rows of A. It follows that the the estimated leverage score $\tilde{l}_j = a_j^T (A_j^T A_j)^{-1} a_j$ is an overestimate of l_j. In a recent paper by [Chhaya et al., 2020], it was shown that the sum of these overestimates can be bounded regardless of how the rows in A are ordered:

Lemma 13 ([Chhaya et al., 2020])**.**

$$\sum_{i=1}^{n} \tilde{l}_j \in O(d + d \log\|A\| - \min_{i \in [n]} \|a_i\|)$$

Next, we show how a simple algorithm that computes $\tilde{l}_j$ in an online manner (passing row by row over the data stream) can be constructed requiring only $\mathcal{O}(d^2)$ of working memory. The idea is to only keep the matrix $A_j^T A_j \in \mathbb{R}^{d \times d}$ in memory and update it for every new row a_{j+1} using a rank one update $A_{j+1}^T A_{j+1} = A_j^T A_j + a_{j+1} \cdot a_{j+1}^T$. See [Golub and van Loan, 2013] for more on matrix multiplication using outer products. The algorithm is given in algorithm 1.

Algorithm 1: Online Leverage Scores

```
Input: Matrix $A \in \mathbb{R}^{n \times d}$
Output: Online leverage scores $\tilde{l}_i$ for all $i \in [n]$
Initialize $M_0 = 0^{d \times d}$
foreach $a_i$ := $i$'th row vector of $A$, $a_i \in \mathbb{R}^d$ do
    $M_i = M_{i-1} + a_i \cdot a_i^T$
    $\tilde{l}_i = a_i^T M_i^{\dagger} a_i$
return $\tilde{l}_i,\ i \in [n]$
```

5 Probit Regression

Situation: We have n data points $(x_i, y_i),\ i = 1, ..., n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.

Probit Model: y_i is a realization of the random variable Y_i. $Y_1, ..., Y_n$ are independent. The distribution of Y_i is as follows:

$$P(Y_i = 1 | x_i; \beta) = \Phi(x_i^T \beta)$$
$$P(Y_i = -1 | x_i; \beta) = 1 - \Phi(x_i^T \beta) = \Phi(-x_i^T \beta)$$

where $\beta \in \mathbb{R}^d$. It follows that

$$P(Y_i = y_i | x_i; \beta) = \Phi(y_i x_i^T \beta)$$

Likelihood: The likelihood of a parameter vector β is given as follows:

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i | x_i; \beta) = \prod_{i=1}^{n} \Phi(y_i x_i^T \beta)$$

The negative log-likelihood that we wish to minimize is:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} \log \Phi(y_i x_i^T \beta)$$

The weighted case: We introduce sample weights $w_i \in \mathbb{R}_{>0}$ comprising a weight vector $w \in \mathbb{R}^n_{>0}$. Further, let $g(z) = -\log \Phi(-z)$. The objective function now becomes:

$$f_w(\beta) = \sum_{i=1}^{n} w_i g(-y_i x_i^T \beta)$$

To make the notation easier, we define $z_i = -y_i x_i^T$ and introduce the matrix $Z \in \mathbb{R}^{n \times d}$ with row vectors $Z_i = z_i$. This gives us:

$$f_w(\beta) = \sum_{i=1}^{n} w_i g(z_i \beta)$$

Gradient: The gradient of the objective function is needed during optimization. To derive it, we first need the derivative of $g(z)$:

$$g'(z) = \frac{d}{dz} - \log \Phi(-z) = \frac{\phi(z)}{\Phi(-z)}$$

Now we can calculate the gradient of the objective function as follows:

$$\frac{\partial f_w(\beta)}{\partial \beta} = \sum_{i=1}^{n} w_i \frac{\partial g(z_i \beta)}{\partial \beta} = \sum_{i=1}^{n} w_i z_i g'(z_i \beta)$$

Lemma 14. *Let $g(z) = -\log \Phi(-z)$. Then it holds for all $z \geq 0$ that:*

$$\frac{1}{2} z^2 \leq g(z)$$

Proof. The following relationship holds for all $z \geq 1$:

$$\begin{aligned}
\Phi(-z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z} \exp\left(-\frac{1}{2}x^2\right) dx \\
&\leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z} -x \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \\
&\leq \exp\left(-\frac{1}{2}z^2\right)
\end{aligned}$$

We therefore have for $z \geq 1$:

$$e^{g(z)} = e^{-\log \Phi(-z)} = \frac{1}{\Phi(-z)} \geq e^{\frac{1}{2}z^2}$$

Since $\exp(\cdot)$ is a monotonically increasing function, it follows that $g(z) \geq \frac{1}{2}z^2$ for all $z \geq 1$.
Let us now turn to the case when $0 \leq z \leq 1$. Both $g(z)$ and $\frac{1}{2}z^2$ are monotonically increasing and continuous functions for $0 \leq z \leq 1$. Together with the fact that $g(0) > \frac{1}{2}$ it follows for all $0 \leq z \leq 1$ that

$$g(z) \geq g(0) > \frac{1}{2} = \max_{0 \leq z \leq 1} \frac{1}{2}z^2 \geq \frac{1}{2}z^2$$

which concludes the proof. □

Lemma 15. *Let $g(z) = -\log \Phi(-z)$. Then it holds for all $z \geq 2$ that:*

$$g(z) \leq z^2$$

Proof. We first show that $\Phi(-z) \geq \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2}$ for all $z \geq 0$. In order to prove this lower bound, we define $h(z) = \Phi(-z) - \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2}$ and show that $h(z)$ is positive for all $z \geq 0$. The derivative $h'(z) = -\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2}z^2}}{(z^2+1)^2}$ is negative for all z, so $h(z)$ is a monotonically decreasing function. Also, it clearly holds that $h(0) > 0$ and $\lim_{z \to \infty} h(z) = 0$. It follows that $h(z) \geq 0$ for all $z > 0$ which proves the lower bound.

In the next step, we use this result to show that $e^{z^2} \cdot \Phi(-z) \geq 1$ for all $z \geq 2$:

$$\begin{aligned} e^{z^2} \cdot \Phi(-z) &\geq e^{z^2} \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2} \\ &= e^{\frac{1}{2}z^2} \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} \\ &= e^{\frac{1}{2}z^2} \frac{1}{\frac{4}{3}(z^2+1)} \frac{\frac{4}{3}z}{\sqrt{2\pi}} \\ &\geq \frac{e^{\frac{1}{2}z^2}}{\frac{4}{3}(z^2+1)} \\ &\geq \frac{e^{\frac{1}{2}z^2}}{e^{\frac{1}{2}z^2}} \\ &= 1 \end{aligned}$$

From this it follows directly that $\frac{1}{\Phi(-z)} \leq e^{z^2}$ and thus we have for all $z \geq 2$:

$$e^{g(z)} = e^{-\log \Phi(-z)} = \frac{1}{\Phi(-z)} \leq e^{z^2}$$

Since $\exp(\cdot)$ is monotonically increasing, the claim that $g(z) \leq z^2$ for all $z \geq 2$ follows as a direct consequence.

The ideas for these proofs are based on the work in [Gordon, 1941]. □

References

[Chhaya et al., 2020] Chhaya, R., Choudhari, J., Dasgupta, A., and Shit, S. (2020). Streaming coresets for symmetric tensor factorization. *CoRR*, abs/2006.01225.

[Cohen et al., 2020] Cohen, M. B., Musco, C., and Pachocki, J. (2020). Online row sampling. *Theory of Computing*, 16(15):1–25.

[Feldman et al., 2020] Feldman, D., Schmidt, M., and Sohler, C. (2020). Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657.

[Golub and van Loan, 2013] Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations*. JHU Press, fourth edition.

[Gordon, 1941] Gordon, R. D. (1941). Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366.

[Kearns and Vazirani, 1994] Kearns, M. J. and Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press.

[Kremer et al., 1999] Kremer, I., Nisan, N., and Ron, D. (1999). On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49.

[Langberg and Schulman, 2010] Langberg, M. and Schulman, L. J. (2010). Universal ϵ-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 598–607.

[Linial et al., 1991] Linial, N., Mansour, Y., and Rivest, R. L. (1991). Results on learnability and the vapnik-chervonenkis dimension. *Information and Computation*, 90(1):33–49.

[Munteanu et al., 2018] Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. (2018). On coresets for logistic regression. In *Advances in Neural Information Processing Systems 31, (NeurIPS)*, pages 6562–6571.