Classification using PGMs

1 Setup and Notation

We are concerned with the classification task of predicting a discrete response $Y \in \mathcal{Y}$ given a set of discrete predictors $X = (X_1, \ldots, X_p) \in \mathcal{X}^p$. For simplicity, we will consider the case where $\mathcal{Y} = \mathcal{X} = \{0, 1\}$, and consider the general discrete case in the sequel. We are given n samples $D = \{(X^{(i)}, Y^{(i)})\}_{i=1}^n$.

The goal is to contrast *generative* and *discriminative* approaches for classification. In a generative model, we learn the joint distribution $P(X, Y)$, and predict the response using the induced conditional distribution $P(Y|X)$. In a discriminative model, we directly learn the conditional distribution $P(Y|X)$.

We are interested in the case where the distribution over the predictors X could be represented using graphical models.

2 Generative Models

We assume that the conditional distribution of X given Y for $Y \in \{0, 1\}$ is distributed as a discrete graphical model. So, there is a separate discrete graphical model associated with $\mathbb{P}(X|Y = 0)$ and with $\mathbb{P}(X|Y = 1)$. In the case of either graphical model, the notation and setting described below will be used.

2.1 Graphical model notation and assumptions.

Let $X = (X_1, X_2, \ldots, X_p)$ denote a random vector, with each variable X_s taking values in a corresponding set $\mathcal{X}_s$. Say we are given an undirected graph G with vertex set $V = \{1, \ldots, p\}$ and edge set E, so that each random variable X_s is associated with a vertex $s \in V$. The pairwise Markov random field associated with the graph G over the random vector X is the family of distributions of X which factorize as $\mathbb{P}(x) \propto \exp\left\{\sum_{(s,t)\in E} \phi_{st}(x_s, x_t)\right\}$, where for each edge $(s, t) \in E$, ϕ_{st} is a mapping from pairs $(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t$ to the real line.[1] For models involving discrete random variables, the pairwise assumption involves no loss of generality, since any Markov random field with higher-order interactions can be converted (by introducing additional variables) to an equivalent Markov random field with purely pairwise interactions (see Wainwright and Jordan [5] for details of this procedure).

Ising Model. In this paper, we focus on the special case of the Ising model in which $X_s \in \{-1, 1\}$ for each vertex $s \in V$, and $\phi_{st}(x_s, x_t) = \theta^*_{st} x_s x_t$ for some parameter $\theta^*_{st} \in \mathbb{R}$, so that the distribution takes the form

$$\mathbb{P}_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp\left\{\sum_{(s,t)\in E} \theta^*_{st} x_s x_t\right\}. \tag{1}$$

The partition function $Z(\theta^*)$ ensures that the distribution sums to one. As we are assuming that there are many conditional independence properties characterizing $\mathbb{P}_{\theta^*}(x)$ θ^* can be considered to be sparse.

[1] Note that, for $(s, t) \notin E$, it is convenient to define $\phi_{st}(x_s, x_t) = 0$.

2.2 Parameters and their estimation.

The graphical models associated with $\mathbb{P}(X|Y=0)$ and with $\mathbb{P}(X|Y=1)$ are distinct. Let $\mathbb{P}(X|Y=0)$ and $\mathbb{P}(X|Y=1)$ be parametrized by different parameter vectors $\theta^{*(0)}$ and $\theta^{*(1)}$ respectively, and let the associated graphs be $G_0=(V_0,E_0)$ and $G_1=(V_1,E_1)$. Let $D_i=\{(x,i)|(x,i)\in D\}$.

2.2.1 Model Estimation

Scheme 1. To estimate $\theta^{*(0)}$ and $\theta^{*(1)}$, one could follow the scheme proposed by Ravikumar et. al. [4] and estimate the p parameters $\theta^{*(i)}_{\setminus r}=\left\{\theta^{*(i)}_{rt}|t\in V\setminus\{r\}\right\}$ associated with each vertex r by solving:

$$\hat{\theta}^{(i)}_{\setminus r}=\arg\min_{\theta_{\setminus r}} l(\theta_{\setminus r}|D_i)+\lambda_i\left\|\theta_{\setminus r}\right\|_1, \tag{2}$$

where $l(\theta_{\setminus r}|D_i)=-\sum_{x\in D_i}\log\mathbb{P}_{\theta_{\setminus r}}(x_r|x_{V\setminus\{r\}})$.
Scheme 2. Suppose that $\theta^{*(0)}$ and $\theta^{*(1)}$ share sparsity structure. Then, one can do feature selection by solving:

$$(\hat{\theta}^{(0)},\hat{\theta}^{(1)})=\arg\min_{\theta^0,\theta^1} l(\theta^1|D_1)+l(\theta^0|D_0)+\lambda\sum_j\left\|[\theta^1_j\ \theta^0_j]\right\|_2. \tag{3}$$

2.2.2 Using the classifier

Once we have estimated the generative model parameters, we can utilize the discriminative classifier described in Equation 4, which results from the computation $\mathbb{P}(Y=1|x)\propto\mathbb{P}(x|Y=1)\mathbb{P}(Y=1)$.

3 Discriminative Model

Consider $\mathbb{P}(y|x)$ induced by $(\theta^{*0},\theta^{*1})$, and let $\mathbb{P}(Y=1)=q$. Let $E=E_0\cup E_1$. $\mathbb{P}(y|x)$ is specified by

$$\mathbb{P}(Y=1|X=x)=\frac{\exp(\theta^*_0+\sum_{(s,t)\in E}\theta^*_{st}x_sx_t)}{1+\exp(\theta^*_0+\sum_{(s,t)\in E}\theta^*_{st}x_sx_t)}, \tag{4}$$

where $\theta^*_0=\log(\frac{q}{Z(\theta^{*1})})-\log(\frac{1-q}{Z(\theta^{*0})})$, and $\theta^*_{st}=\theta^{*1}_{st}-\theta^{*0}_{st}$.

Given D, we could estimate this discriminative model directly, without first estimating θ^{*1} and θ^{*0}, by solving:

$$\hat{\theta}=\arg\min_{\theta} l(\theta|D)+\lambda\sum_{j>0}|\theta_j|, \tag{5}$$

where $l(\theta|D)=-n^{-1}\sum_{(x,y)\in D}\log\mathbb{P}_\theta(y|x)$. By looking at the sparsity pattern in θ^*, we estimate the edge-set $\hat{E}$.

Note that the discriminative classifier described by Equation 4 corresponds to using the halfspace

$$f(x,\theta^{*i})=\theta^{*i}_0+\sum_{(s,t)\in E}\theta^{*i}_{st}x_sx_t\geq 0 \tag{6}$$

for classification.

4 Bounds on Classification Error

We wish to analyze the expected risk of the classifiers above as a function of n and p. In doing this, we would be following the footsteps of Ng and Jordan [2].

The hypothesis class considered while estimating θ^* directly by solving Problem 5 is at least as large as the hypothesis class considered while estimating θ^* by first estimating the generative model parameters. Hence, we note that as $n \to \infty$ the expected risk of the classifier in the former case is at least as low as the estimated risk of the classifier learned by estimating θ^{*1} and θ^{*2} first. We now analyze the n required to learn a classifier which is almost as the classifiers mentioned for the $n \to \infty$ case.

4.1 Discriminative case

4.1.1 Loose bounds on sample complexity

As, in the case of logistic regression, the 0/1 misclassification error $\epsilon^m(\theta)$ is related to the logloss $\epsilon^l(\theta)$ by $\epsilon^l(\theta) \geq (\log 2)\epsilon^m(\theta)$, following the analysis in [3], we note that $n = \Omega(\log p)poly(|E|)$ suffices. [2]

4.1.2 Tighter bounds

Theorem 1. *Suppose that θ^* is estimated as $\hat{\theta}$ as in Equation 5, with the exception that $|\theta_0^*|$ is included in the regularization.* [3]

Let $f(x, \theta^*)$ be the linear discriminant corresponding to Equation 6, and let $sgn(f(x, \theta^*))$ be the corresponding classifier. Similarly, let $sgn(f(x, \theta))$ be the classifier corresponding to $\hat{\theta}$.

Suppose that θ^* also satisfies the conditions specified for the application of Theorem 5 of [1] in the below proof. In addition, suppose that $Pr_x(f(x, \theta^*) \leq \mu) \leq p_2$ for $\mu \geq k|E|\sqrt{\log p}/n$ for some constant k specified in the proof.

Then, $Pr_y(sgn(f(x, \theta^*)) \neq sgn(f(x, \theta))) \leq p_1 + p_2$ for the failure probability p_1 described in the statement of Theorem 5 in [1] [4].

Proof. First, we rewrite the optimization problem so that Theorem 5 from [1] can be applied.

Feature map. Any data point x is mapped to z, a vector whose components are defined by: $z_0 = 1$ and $\{z_{ij} = x_i x_j | \forall i \neq j\}$. The length of z is given by $p' = \binom{p}{2} + 1 = \Theta(p^2)$. Using this feature map, the set of points D is mapped to the set of points D'. Once this is done, all z is normalized, so that $\forall i : |z_i| = 1/\sqrt{n}$.

Equivalent problem. Next, the define parameters $\theta' = \sqrt{n}\theta^*$. Thus, equivalent to the model described in Equation 4, we have the model where $\mathbb{P}(Y = i) = \sigma(i\langle\theta', z\rangle)$, which is estimated by solving the optimization problem:

$$\hat{\theta}' = \arg\min_{\theta} l(\theta|D') + \lambda \sum_{j \geq 0} |\theta_j|, \tag{7}$$

where $l()$ is defined as in Equation 5. Solving this optimization problem is equivalent to solving the optimization problem specified in the theorem statement.

Applying Theorem 5 from [1], using $\lambda = k\sqrt{\frac{\log p}{n}}$ for some constant k, we then conclude that, with probability at least $1 - p_1$:

$$\left\|\hat{\theta}' - \theta'\right\|_1 = \sqrt{n}\left\|\hat{\theta} - \theta^*\right\|_1 \leq k_1|E|\sqrt{\frac{\log p}{n}},$$

where k_1 is another constant.

As $\|z\|_\infty = 1$, we can conclude that

$$|f(x, \theta^*) - f(x, \hat{\theta})| \leq k_1|E|\sqrt{\log p}/n.$$

Applying Claim 2, we have the result. □

[2] It is possible that actually $n = \Omega(|E| \log p)$ suffices when we separate the feature selection step from the final estimation step and use the VC dimension for linear classifiers: **[Check]**. This could improve known sample complexity for ℓ_1 regularized logistic regression.

[3] Note that using [1] seems to require including θ_0 in the regularization.

[4] The proof of this theorem is yet to be verified and fixed to suit our purposes.

Remark 1. *Note that in Theorem 1, we considered an optimization problem which was slightly different from Equation 5.*

To achieve misclassification rate which is at most $p_1 + p_2$ greater than the best achievable misclassification rate, number of samples required is given by:

$$n = O(|E|\frac{\sqrt{\log p}}{\mu}).$$

Note that p_1 decreases exponentially with n, so it suffices to focus on controlling p_2 and μ. In particular, if we impose conditions on θ^{*i} such that $\mu = \Omega(|E|)$, we have:

$$n = O(\sqrt{\log p}).$$

4.2 Generative case

4.2.1 Loose bounds on sample complexity

Given that feature selection is done and we have an estimate $\hat{E}$ of E, using the VC dimension of linear classifiers, we observe that $n = O(|\hat{E}|)$ examples are sufficient to achieve a low classification error rate. Below we consider ways of characterizing the number of samples n required to get estimate $\hat{E} \supset E$ which is not too much larger than E.

Consider Scheme 1. From [4], we know that if θ^{*1} and θ^{*2} satisfy certain strong conditions, $n = O(d^3 \log p)$ is sufficient to estimate E_i accurately. Considering the fact that even $\hat{E}_i \supset E_i$ suffices as long as it is not too much bigger than E_i, we can probably do better: that is, we should be able to use more relaxed conditions on θ^{*i}, and we can make do with smaller n.

4.2.2 Tighter bounds for Scheme 1

Claim 1. *Suppose that θ^{*i} are estimated using Scheme 1. Also suppose that θ^{*i} satisfy the conditions required by Theorem 1 in [4]. Then, with probability at least $1 - p_1$ for the failure probability p_1 described in the statement of Theorem 1 in [4],*

$$\left\|\theta^{*i} - \hat{\theta}^i\right\|_2 \leq k_1 \sqrt{min(p, |E_i|)} \sqrt{\frac{d \log p}{n}}$$

and

$$\left\|\theta^{*i} - \hat{\theta}^i\right\|_\infty \leq k_2 \sqrt{\frac{d \log p}{n}},$$

for some constants k_1 and k_2.

Proof. Considering the proof of Proposition 1 on page 17 of [4], and taking $\lambda_n = k\sqrt{\frac{\log p}{n}}$, we have:

$$\left\|\theta^{*i}_{\backslash r} - \hat{\theta}^i_{\backslash r}\right\|_\infty \leq \left\|\theta^{*i}_{\backslash r} - \hat{\theta}^i_{\backslash r}\right\|_2 \leq c\sqrt{\frac{d \log p}{n}},$$

where c some constant independent of d, p, n and $\theta_{\backslash r}$ is the vector of parameters $\{\theta_{r,j},\ \forall j \in V \backslash \{r\}\}$.

From this, we already have the bound on $\left\|\theta^{*i} - \hat{\theta}^i\right\|_\infty$.

We have the bound $\left\|\theta^{*i} - \hat{\theta}^i\right\|_2 \leq c\sqrt{p}\sqrt{\frac{d \log p}{n}}$ by applying the generalized Pythagoras theorem. We also observe from the above that $\left\|\theta^{*i}_{j,k} - \hat{\theta}^i_{j,k}\right\|_2 \leq c\sqrt{\frac{d \log p}{n}} \forall (i,j) \in E$. Again, applying the generalized Pythagoras theorem, we have: $\left\|\theta^{*i} - \hat{\theta}^i\right\|_2 \leq c\sqrt{E}\sqrt{\frac{d \log p}{n}}$. Combining these, we have the bound on $\left\|\theta^{*i} - \hat{\theta}^i\right\|_2$. □

Theorem 2. *Suppose that θ^{*i} are estimated as $\hat{\theta}^{(i)}$ using Scheme 1, and let $f(x, \theta^*)$ be the linear discriminant corresponding to Equation 6, and let $sgn(f(x, \theta^*))$ be the corresponding classifier. Similarly, let $sgn(f(x, \theta))$ be the classifier corresponding to $\hat{\theta}^{(i)}$.*

Suppose that θ^{*i} satisfy the conditions described in Claim 1. In addition, suppose that $Pr_x(f(x, \theta^*) \leq \mu) \leq p_2$ for $\mu \geq k\sqrt{|E|}\sqrt{\min(p, |E|)}\sqrt{\frac{d \log p}{n}}$ for some constant k specified in the proof.

Then, $Pr_y(sgn(f(x, \theta^*)) \neq sgn(f(x, \theta))) \leq p_1 + p_2 + p_3$ for the failure probability p_1 described in the statement of Theorem 1 in [4] and p_3 is specified in the proof below.

Proof. For some constant k_1 from Claim 1, with probability at least $1 - p_1$,

$$\|\theta - \theta^*\|_2 \leq k_1 \sqrt{min(p, |E|)} \sqrt{\frac{d \log p}{n}}.$$

Given that this holds, as $|x_s x_t| = 1$,

$$\sum_{(s,t) \in E} (\theta_{st} - \theta^*_{st}) x_s x_t \leq \|\theta - \theta^*\|_2 \sqrt{|E|},$$

from applying the Holder inequality.

Hence, using Claim 3

$$\log(\frac{Z(\theta^{*0})}{Z(\hat{\theta}^{(0)})}) - \log(\frac{Z(\theta^{*1})}{Z(\hat{\theta}^{(1)})}) = O(\|\theta - \theta^*\|_2 \sqrt{|E|}).$$

$$\theta_0 - \theta^*_0 = \log(\frac{q}{\hat{q}}) - \log(\frac{1-q}{1-\hat{q}}) + \log(\frac{Z(\theta^{*0})}{Z(\hat{\theta}^{(0)})}) - \log(\frac{Z(\theta^{*1})}{Z(\hat{\theta}^{(1)})}),$$

where $q = Pr(Y = 1)$ and $\hat{q}$ is its estimate using n samples. Using Chernoff bounds, we know that $|q - \hat{q}| = O(\frac{1}{\sqrt{n}})$ with probability at least $1 - p_3$. So, we claim that

$$|\log(\frac{q}{\hat{q}}) - \log(\frac{1-q}{1-\hat{q}})| = O(\frac{1}{\sqrt{n}}).$$

Hence, for some constant k_2:

$$|\theta_0 - \theta^*_0| \leq k_2 \sqrt{|E|} \sqrt{min(p, |E|)} \sqrt{\frac{d \log p}{n}}$$

.

Applying these bounds to discriminant functions described by Equation 6, we get

$$|f(x, \theta^*) - f(x, \theta)| \leq k \sqrt{|E|} \sqrt{min(p, |E|)} \sqrt{\frac{d \log p}{n}}.$$

Applying Claim 2, we have the result. □

Remark 2. *Consider Theorem 2. To achieve misclassification rate which is at most $p_1 + p_2 + p_3$ greater than the best achievable misclassification rate, number of samples required is given by:*

$$n = O(|E| \min(|E|, p) \frac{d \log p}{\mu^2}).$$

Note that p_1 and p_3 decrease exponentially with n, so it suffices to focus on controlling p_2 and μ. In particular, if we impose conditions on θ^{*i} such that $\mu = \Omega(|E|)$, we have:

$$n = O(d \log p).$$

4.3 Technical theorems

Claim 2. *Suppose that a family of binary classifiers is defined by* $sgn(f(x))$*, where* $f : R^p \to R$ *and* $sgn(x)$ *is the sign function. Consider two classifiers* $sgn(f(x))$ *and* $sgn(g(x))$*.*

Then, if $Pr_x(|f(x)| < \mu) \leq p_1$ *and* $Pr_x(|g(x) - f(x)| \geq \mu) \leq p_2$*. Then,* $Pr_x(sgn(f(x)) \neq sgn(g(x))) \leq p_1 + p_2$*.*

Claim 3. *If* $|a_i - b_i| \leq \epsilon$*, then* $\log(\frac{\sum_i exp(a_i)}{\sum_i exp(b_i)}) \leq O(\epsilon)$*.*

Proof. Suppose that $|a_i - b_i| \leq \epsilon$. Then, $|exp(a_i) - exp(b_i)| \leq exp(a_i)(1 - exp(\epsilon)) = exp(a_i)O(\epsilon)$ using the McLaurin series for $exp(\epsilon)$.

$\frac{\sum_i exp(a_i)}{\sum_i exp(b_i)} \leq 1 + (\frac{|\sum_i exp(a_i) - \sum_i exp(b_i)|}{\sum_i exp(b_i)}) \leq 1 + (\frac{\sum_i exp(a_i)}{\sum_i exp(b_i)})O(\epsilon)$.

So, $\frac{\sum_i exp(a_i)}{\sum_i exp(b_i)} \leq \frac{1}{1-O(\epsilon)} \leq 1 + O(\epsilon)$.

So, $\log(\frac{\sum_i exp(a_i)}{\sum_i exp(b_i)}) \leq \log(1 + O(\epsilon)) \leq O(\epsilon)$, using the McLaurin series for $\log(1 + x)$. □

References

[1] Francis Bach. Self-concordant analysis for logistic regression. *CoRR*, abs/0910.4627, 2009. informal publication.

[2] A. Ng and M. Jordan. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, 2001.

[3] A. Y. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *International Conference on Machine Learning*, 2004.

[4] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[5] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics, September 2003.