The Asymptotics of Semi-Supervised Learning in Discriminative Probabilistic Models

Nataliya Sokolovska, Olivier Cappé
& François Yvon

TELECOM ParisTech & CNRS

Université Paris-Sud & CNRS

IMCL 2008

Outline

Semi-Supervised Classification

Refers to the use of some prior knowledge about the marginal distribution of the features $\{X_i\}$ to improve supervised classification (prediction of the label Y_i from X_i)

Why Is This Topic Important?

- In many learning applications, unlabeled data is plentiful and can be collected at almost no cost, whereas labeled data is comparatively rare and more costly to gather
- Semi-supervised learning is related but very different from the statistical issue of *missing data*, where only a few labels would be unobserved

(Subjective) Literature Survey

Several (classifier-dependent) approaches based on the general intuition that semi-supervised learning is useful mostly in face of low Bayes error

This cluster assumption is generally implemented by adding a penalty term to the supervised learning criterion so as to

- force decisions boundaries to cross only low density regions
- make decisions as unambiguous as possible in high density regions

Some Concerns

- Improvement over the situation where nothing is known about the marginal $q(x)$ not always guaranteed in practice (!)
- Generative and discriminative probabilistic models appear to behave very differently in this context

This Contribution

Assumes

- a discriminative probabilistic model $g(y|x;\theta)$
- that the marginal $q(x)$ is fully known
- that the feature $\mathcal{X}$ and label sets $\mathcal{Y}$ are finite

and provides an asymptotically (i.e., $n \to \infty$) optimal semi-supervised estimation procedure and discusses its performance (compared to that of usual supervised learning)

Notations

- X_i features
- Y_i labels
- n training sample size
- $\pi(x, y)$ joint
- $\eta(y|x)$ conditional
- $q(x)$, $p(y)$ marginals
- $g(y|x;\theta)$ model conditional, with parameter $\theta \in \Theta$, and

$$\ell(y|x;\theta) = -\log g(y|x;\theta)$$

is the negated conditional log-likelihood

Stratified Sampling

Well-known Principle in Survey Sampling

- In a two-way contingency table, the maximum likelihood estimate of the joint cell probability $\pi(x, y)$ when the marginal $q(x)$ is known is given by

$$\hat{\pi}_n^s(x, y) = \frac{\sum_{i=1}^n \mathbb{1}\{X_i = x, Y_i = y\}}{\sum_{j=1}^n \mathbb{1}\{X_j = x\}} q(x)$$

- Its asymptotic variance is

$$v_n^s(x, y) = \pi(x, y)(1 - \eta(y|x))$$

compared to $\pi(x, y)(1 - \pi(x, y))$ for the un-stratified estimator $\hat{\pi}_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i = x, Y_i = y\}$

Stratified Sampling, Contd.

The classical statistical use of this result consists in estimating marginal probabilities $p(y)$ by

$$\hat{p}^s_n(y) = \sum_x \hat{\pi}^s_n(x, y)$$

The stratified estimator $\hat{p}^s_n(y)$ has asymptotic variance

$$\sum_x q(x)\eta(y|x)(1 - \eta(y|x)) = \mathrm{E}_q \left(\mathrm{V}_\eta \left[\mathbb{1}\{Y = y\} \middle| X \right] \right)$$

which is smaller than $\mathrm{V}_\pi \left[\mathbb{1}\{Y = y\} \right]$ for the un-stratified estimator $\hat{p}_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i = y\}$

The difference between both asymptotic variances may be expressed as $\mathrm{V}_q(\mathrm{P}_\eta[Y = y|X])$ (due to the Rao-Blackwell variance decomposition)

The Performance Criterion

Let $g(y|x;\theta)$ denote the conditional probability associated with a discriminative probabilistic classifier; we consider log-likelihood-based methods that aim at minimizing the logarithmic risk

$$r(\theta) = \mathrm{E}_\pi[\ell(Y|X;\theta)]$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$

Estimates $\hat{\theta}_n$ obtained with such methods typically satisfy

- $\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{L} \mathcal{N}(0, \Sigma(\theta_\star))$, where $\theta_\star = \arg\min_{\theta\in\Theta} r(\theta)$
- $n(\mathrm{E}_{\pi^{\otimes n}} r(\hat{\theta}_n) - r(\theta_\star)) \to \frac{1}{2}\,\mathrm{trace}(J(\theta_\star)\Sigma(\theta_\star))$, where

$$J(\theta_\star) = \mathrm{E}_\pi\left[\nabla_{\theta^{\mathrm{T}}}\nabla_\theta \ell(Y|X;\theta_\star)\right]$$

The Performance Criterion

Let $g(y|x;\theta)$ denote the conditional probability associated with a discriminative probabilistic classifier; we consider log-likelihood-based methods that aim at minimizing the logarithmic risk

$$r(\theta) = \mathrm{E}_\pi[\ell(Y|X;\theta)]$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$

Estimates $\hat{\theta}_n$ obtained with such methods typically satisfy

- $\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{L} \mathcal{N}(0, \Sigma(\theta_\star))$, where $\theta_\star = \arg\min_{\theta\in\Theta} r(\theta)$
- $n(\mathrm{E}_{\pi^{\otimes n}} r(\hat{\theta}_n) - r(\theta_\star)) \to \frac{1}{2}\,\mathrm{trace}(J(\theta_\star)\Sigma(\theta_\star))$, where

$$J(\theta_\star) = \mathrm{E}_\pi\left[\nabla_{\theta^{\mathrm{T}}}\nabla_\theta \ell(Y|X;\theta_\star)\right]$$

If the model is well-specified, that is $\eta(y|x) = g_{\theta_\star}(y|x)$, not only is r minimal at $\theta_\star$ but one has the stronger property that $\mathrm{E}_\eta[\ell_\theta(Y|X)|X=x]$ is minimized at $\theta_\star$, for all values of x

An Asymptotically Optimal Estimator

Our Main Result

We propose the following weighted maximum-likelihood estimator

$$\hat{\theta}_n^s = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}} \ell(Y_i | X_i; \theta)$$

An Asymptotically Optimal Estimator

Our Main Result

We propose the following weighted maximum-likelihood estimator

$$\hat{\theta}_n^s = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}} \ell(Y_i|X_i;\theta)$$

which achieves the minimal asymptotic variance of $\Sigma(\theta_\star) = J^{-1}(\theta_\star)H(\theta_\star)J^{-1}(\theta_\star)$, where

$$H(\theta_\star) = \mathrm{E}_q\left(\mathrm{V}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X\right]\right)$$

For comparison, the unweighted maximum-likelihood estimator $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i|X_i;\theta)$ has asymptotic variance $J^{-1}(\theta_\star)I(\theta_\star)J^{-1}(\theta_\star)$, where

$$I(\theta_\star) = \mathrm{V}_\pi\left[\nabla_\theta \ell(Y|X;\theta_\star)\right]$$

Main Proof Argument

$$\hat{\theta}_n^s = \arg\min_{\theta \in \Theta} \sum_x \sum_y \hat{\pi}_n^s(x,y)\ell(y|x;\theta)$$

where

$$\hat{\pi}_n^s(x,y) = \begin{cases} \frac{\sum_{i=1}^n \mathbb{1}\{X_i=x, Y_i=y\}}{\sum_{j=1}^n \mathbb{1}\{X_j=x\}} q(x) & \text{if } \sum_{i=1}^n \mathbb{1}\{X_i = x\} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Hence, $\hat{\theta}_n^s$ is the maximum likelihood estimate of θ under the constraint that $\sum_y \pi(x,y) = q(x)$ $\square$

In the Case of Binary Logistic Regression

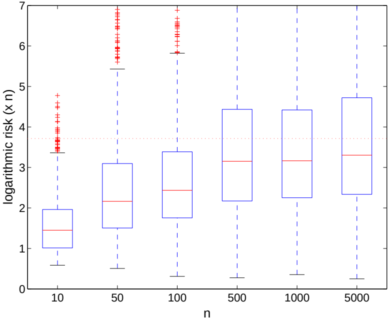
$$
\begin{aligned}
J(\theta_\star) &= \mathrm{E}_q\left[g(1|X;\theta_\star)\{1-g(1|X;\theta_\star)\}XX^{\mathrm{T}}\right] \\
H(\theta_\star) &= \mathrm{E}_q\left[\eta(1|X)(1-\eta(1|X))XX^{\mathrm{T}}\right] \\
I(\theta_\star) &= \mathrm{E}_q\big[\{\eta(1|X)(1-\eta(1|X)) \\
&\qquad\qquad + (\eta(1|X)-g(1|X;\theta_\star))^2\}XX^{\mathrm{T}}\big]
\end{aligned}
$$

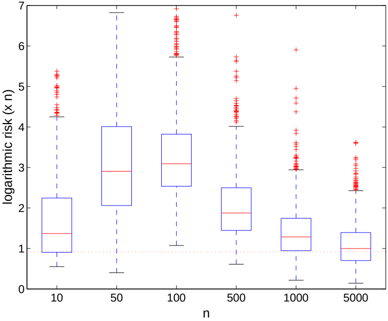
- The difference

$$
I(\theta_\star) - H(\theta_\star) = \mathrm{E}_q\left[\{\eta(1|X) - g(1|X;\theta_\star)\}^2 XX^{\mathrm{T}}\right]
$$

  is all the more significant that the fit achievable by the model is poor
- The matrix $H(\theta_\star)$ may be significantly smaller than $I(\theta_\star)$ only when the Bayes error associated with π is small

Simulation Experiment With Binary Logistic Regression





Boxplots of the scaled excess logarithmic risk as a function of the number of observations. Left: for logistic regression, $n(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$; right: for the semi-supervised estimator, $n(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}^s_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$ (Bayes error 1.7%, model error 9.4%).

Connection With the Covariate Shift Problem

Covariate Shift

Assuming a classifier trained from data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i is marginally distributed according to $q_0(x)$; how to adapt the classifier when the future X_is are distributed under $q_1(x) \neq q_0(x)$?

- If q_1 is known, weights in the proposed semi-supervised estimator (used with $q = q_1$) are asymptotically equivalent to $\frac{1}{n}\frac{q_1}{q_0}(X_i)$ and the algorithm converges to

$$\theta_{1\star} = \arg\min_{\theta \in \Theta} \mathrm{E}_{\pi_1}[\ell(Y|X;\theta)]$$

- The associated asymptotic covariance matrix is smaller than that of the importance ratio weighted estimator (!)

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \frac{q_1}{q_0}(X_i)\ell(Y_i|X_i;\theta)$$

(which, in addition, assumes knowledge of q_0)

Applications to Larger Scale Problems

When dealing with larger scale problems (see text classification example discussed in the paper), it is no more reasonable to assume that $\mathcal{X}$ is finite

We propose a strategy based on clustering

How To "Estimate $q(x)$"?

The complete unlabeled collection of features is clustered into k clusters, and in the weight expression

$$\frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}}$$

the numerator is replaced by the empirical frequency of the cluster to which X_i belongs while the denominator is replaced by the number of training documents belonging to the same cluster as X_i

Some Conclusions

We have analyzed a simple asymptotically optimal semi-supervised estimation strategy

- the performance analysis provides interesting insight into the potentials of semi-supervised learning in discriminative probabilistic models
- the proposed estimator is easy to implement and also appears to be useful in the covariate-shift scenario
- some ideas on how to generalize the approach to more general settings
- observed practical improvements are sometimes limited but this is, to some extent, predicted by the performance analysis
- the situation where, both, the Bayes error is very low and the number of training samples is very small deserves some attention