Score Function Features for Discriminative Learning: Matrix and Tensor Framework

Janzamin, Sedghi, Anandkumar

Arthur Gretton's notes

February 26, 2015

Public service announcement

- The Journal of Basic and Applied Social Psychology has banned the use of p-values.

Public service announcement

- The Journal of Basic and Applied Social Psychology has banned the use of p-values.
- "We hope and anticipate that banning the null hypothesis significance testing procedure (NHSTP) will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking."

Public service announcement

- The Journal of Basic and Applied Social Psychology has banned the use of p-values.
- “We hope and anticipate that banning the null hypothesis significance testing procedure (NHSTP) will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking.”
- What about Bayesian error analysis?

Public service announcement

- The Journal of Basic and Applied Social Psychology has banned the use of p-values.
- "We hope and anticipate that banning the null hypothesis significance testing procedure (NHSTP) will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking."
- What about Bayesian error analysis?
- "The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist"
- "With respect to Bayesian procedures, we reserve the right to make case-by-case judgments, and thus Bayesian procedures are neither required nor banned from BASP." (an uninformed prior?)

What the paper is about

How do we easily obtain good features for classification?

- Features (expected high order derivatives) of the conditional mean of the output given the input (useful for classification, where this conditional mean is all that matters).
- Such derivatives can be estimated using scores, which come from unlabaled data.

This paper conjectures a set of informative features of these derivatives (with zero evidence).

What the paper is about

How do we easily obtain good features for classification?

- Features (expected high order derivatives) of the conditional mean of the output given the input (useful for classification, where this conditional mean is all that matters).
- Such derivatives can be estimated using scores, which come from unlabaled data.

This paper conjectures a set of informative features of these derivatives (with zero evidence).

Outline:

- How do score functions determine features of the conditional distribution?
- How do we extract features from these scores?

Problem setting

Conditional mean of y (binary) given x:

$$G(x) = \mathbb{E}(y|x).$$

Some useful features for classification might derive from

$$\mathbb{E}\left(\nabla_x^{(m)} G(x)\right),$$

e.g. for $m \leq 3$ (up to third order tensor).
These are hard to compute, and in any case require labeled data.

Simplest case: first order score

Idea: estimate

$$-\nabla \log p(x)$$

Then

$$-\mathbb{E}\left(y \nabla \log p(x)\right) = \mathbb{E}\left(\nabla_x G(x)\right)$$

You can learn $-\nabla \log p(x)$ from unlabeled data, then apply it to many prediction problems.

Next:

- Proof of the above result
- Learning problem which allows us to estimate $-\nabla \log p(x)$.

Simplest case: first order score

Result:

$$-\mathbb{E}\left(y\nabla \log p(x)\right) = \mathbb{E}\left(\nabla_x G(x)\right)$$

Proof in 1-D (from Stein et al., 2004, Proposition 4)
Definitions and conditions:

- Interval $I := [a, b]$ where $-\infty \leq a < b \leq \infty$.
- $p(x)$ a density on I with a regular derivative $p'(x)$ (countably many sign changes, continuous at sign changes)
- Score is

$$\psi(x) = \frac{p'(x)}{p(x)} = \frac{d}{dx} \log p(x)$$

- $G(x) \in \mathcal{F}$ is class of functions where the following integrals exist:

$$\mathbb{E}\left[\left|G'(x)\right|\right] < \infty \qquad \mathbb{E}\left[\left|G(x)\psi(x)\right|\right] < \infty$$

Simplest case: first order score

Proof (continued): Integration by parts:

$$\begin{aligned}\mathbb{E}G'(X) &= \int_I G'(x)p(x)dx \\ &= G(b-)p(b-) - G(a+)p(a+) - \int_I G(x)p'(x)dx \\ &= G(b-)p(b-) - G(a+)p(a+) - \int_I G(x)\psi(x)p(x)dx.\end{aligned}$$

Finally, assuming everything goes to zero at boundaries,

$$\mathbb{E}G'(X) = -\mathbb{E}\left[\underbrace{\mathbb{E}(y|x)}_{G(X)}\psi(x)\right] = -\mathbb{E}\left[y\psi(x)\right].$$

How to learn first order score

One idea is score matching (Hyvarinen, 2005).
Given a parametric model q_θ parametrized by θ,

$$D_F(p, q_\theta) = \int_x p(x) \left\| \frac{\nabla_x p(x)}{p(x)} - \frac{\nabla_x q_\theta(x)}{q_\theta(x)} \right\| dx.$$

How to learn first order score

One idea is score matching (Hyvarinen, 2005).
Given a parametric model q_θ parametrized by θ,

$$D_F(p, q_\theta) = \int_x p(x) \left\| \frac{\nabla_x p(x)}{p(x)} - \frac{\nabla_x q_\theta(x)}{q_\theta(x)} \right\| dx.$$

Again integrating by parts, we get

$$D_F(p, q_\theta) = \int_x p(x) \left(\underbrace{\|\nabla \log p(x)\|^2}_{\text{indep of } \theta} + \|\nabla \log q_\theta(x)\|^2 + 2\Delta \log q_\theta(x) \right) dx$$

where

$$\Delta := \sum_{i \in [d]} \frac{\partial^2}{\partial x_i^2}.$$

Empirically: replace expectation over $p(x)$ with empirical expectation, solve for θ (we do this in infinite exp. family paper)

Another estimate of score functions

From Alain and Bengio (2014): denoising autoencoder is:

$$\mathcal{L}_{\mathrm{DAE}} := \mathbb{E}\left[\ell(x, r(N(x)))\right]$$

where

- $r(N(x))$ is the reconstructed version of x from $N(x)$, $r = g(f(x))$, where f is an encoder, and g is a decoder.
- ℓ is the squared loss, $\ell(x, y) = (x - y)^2$.
- $N(x) = x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

The optimal $r_\sigma*$ takes the form (assuming f, g have the capacity to represent it...)

$$r_\sigma^*(x) = \frac{\mathbb{E}_\epsilon[p(x-\epsilon)(x-\epsilon)]}{\mathbb{E}_\epsilon[p(x-\epsilon)]}$$

and

$$r_\sigma^* = x + \sigma^2 \frac{\partial \log p(x)}{\partial x} + o(\sigma^2) \quad \sigma \to 0.$$

I.e. use denoising autoencoders to get score estimates.

Does this generalize to higher order?

The multivariate score relation:

$$\mathbb{E}\left[\nabla^{(m)} G(x)\right] = \mathbb{E}[G(x) S_m(x)],$$

where the scores

$$S_m(x) = (-1)^m \frac{\nabla_x^{(m)} p(x)}{p(x)}$$

may be defined by recursion,

$$S_m = -S_{m-1}(x) \otimes \nabla_x \log p(x) - \nabla_x S_{m-1}(x).$$

Does this generalize to higher order?

The multivariate score relation:

$$\mathbb{E}\left[\nabla^{(m)} G(x)\right] = \mathbb{E}[G(x) S_m(x)],$$

where the scores

$$S_m(x) = (-1)^m \frac{\nabla_x^{(m)} p(x)}{p(x)}$$

may be defined by recursion,

$$S_m = -S_{m-1}(x) \otimes \nabla_x \log p(x) - \nabla_x S_{m-1}(x).$$

Gaussian case: $p(x) = \frac{1}{\left(\sqrt{2\pi}\right)^{d_x}} e^{-\|x\|^2/2}$. Then $\nabla_x \log p(x) = -x$, and we recover Stein's lemma,

$$\mathbb{E}\left[x G(x)\right] = \mathbb{E}\left[\nabla_x G(x)\right].$$

What features can we get?

Idea: we want features of expected high order derivatives of

$$T := \mathbb{E}\left[\nabla^{(m)} G(x)\right]$$

A tensor has CP-rank k if it can be written as the sum of k rank-1 tensors,

$$T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i.$$

How do we find such a decomposition?

What features can we get?

Algorithm 1 Tensor decomposition via tensor power iteration (Anandkumar et al., 2014b)

Require: 1) Rank-k tensor $T = \sum_{j\in[k]} u_j \otimes u_j \otimes u_j \in \mathbb{R}^{d\times d\times d}$, 2) L initialization vectors $\hat{u}_\tau^{(1)}$, $\tau \in [L]$, 3) number of iterations N.

for $\tau = 1$ **to** L **do**

  for $t = 1$ **to** N **do**

    Tensor power updates (see (15) for the definition of the multilinear form):

$$\hat{u}_\tau^{(t+1)} = \frac{T\left(I, \hat{u}_\tau^{(t)}, \hat{u}_\tau^{(t)}\right)}{\left\| T\left(I, \hat{u}_\tau^{(t)}, \hat{u}_\tau^{(t)}\right)\right\|}, \tag{13}$$

  end for

end for

return the cluster centers of set $\left\{\hat{u}_\tau^{(N+1)} : \tau \in [L]\right\}$ (by Procedure 2) as estimates u_j.

We used the multilinear form

$$T(I, v, w) = \sum_{j,l\in[d]} v_j w_l T(:, j, l).$$