Optimization of Bregman Divergences

Ragib Zaman

An essay submitted in partial fulfillment of
the requirements for the course
COMP6470 - Special Topics in Computing

College of Engineering and Computer Science
Australian National University

December 2018

Abstract

In this report we survey Bregman Divergences and motivate their study by forming connections to areas such as convex optimization, machine learning and statistics. We give particular focus to developing an understanding of the optimization of Bregman Divergences and the connection to Exponential Families and Logistic Regression. In the second chapter we see that the Kullback-Liebler divergence between two members of an exponential family can be expressed as a Bregman Divergence between the distribution parameters. The third chapter gives a brief overview of Generalized Linear Models and Logistic Regression. We give a moderately simplified proof of a theorem of A. Banerjee ([**1**]) which states that the log-odds ratio of the class posteriors is affine (which is the modelling assumption of Logistic Regression) if and only if the class conditional distributions belong to a fixed natural exponential family. In the final chapter, we show how the problem of finding the optimal parameters for the log loss of Logistic Regression can be cast as a Bregman Divergence optimization problem, and prove the convergence of an algorithm for finding those parameters using methods standard in the optimization of Bregman divergences. By simplifying their assumptions and/or proofs where we were able to, we believe we have given a clearer presentation of the results which we have drawn from our references. We hope that the reader will come away with a keen interest in Bregman Divergences and be aware of potential applications in their fields of study.

Acknowledgments

First and foremost I would like to thank Dr. Cheng Soon Ong. I greatly value the wisdom he has shared with me, helping me identify the core ideas in machine learning and the links between. His willingness to give his time so generously and with such patience has been very much appreciated.

I would also like to thank Dr. Weifa Liang and Dr. Stephen Gould for taking the responsibilities of course convener and examiner respectively. Without their contribution I would not have been able to have the pleasure of completing this course.

CONTENTS

CHAPTER 1

Bregman Divergences

1.1. Introduction

The notion of convexity and the properties of convex functions have proven to be very useful, fundamentally due to them being the largest class of function for which we can develop a satisfactory theory of optimization for. As such, convex functions appear naturally as objects of interest in any field which utilizes optimization, such as machine learning, statistics, control theory, quantitative finance and many more.

Bregman divergences encode more precise information about convexity. Suppose that $\Delta \subseteq \mathbb{R}^m$ is a convex set and $F : \Delta \longrightarrow \mathbb{R}$ is continuously differentiable. Then F is a convex function if and only if $F(p) - F(q) - \langle \nabla F(q), p - q \rangle \geq 0$ for all $p, q \in \Delta$. This leads to the definition of Bregman Divergences.

Definition 1.1. Let $F : \Delta \to \mathbb{R}$ be a strictly convex function, continuously differentiable on a closed convex set $\Delta \subseteq \mathbb{R}^m$. The *Bregman Divergence generated by* F is the function $B_F : \Delta \times \Delta \to \mathbb{R}$ defined by

$$B_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle$$

We have restricted the generators of Bregman Divergences to be strictly convex functions so as to be consistent with most other sources. This restriction allows Bregman Divergences to satisfy certain additional properties which we will soon discuss. It is important to note however that some of the properties we discuss do not require the convexity of F, and interesting results can be obtained by considering Bregman Divergences generated by more general F (for example, see [**10**]).

Example 1.2. Two familiar examples are:

- Squared Mahalanobis Distance
  Let $F : \mathbb{R}^m \to \mathbb{R}, x \mapsto x^T Q x$, where Q is a positive definite matrix. Then $B_F(p, q) = (p - q)^T Q (p - q)$. When $Q = I_m$, this is the squared Euclidean distance.
- Kullback-Leibler Divergence (also known as Relative Entropy)
  Let $F : (\mathbb{R}_{>0})^m \to \mathbb{R}, x \mapsto \sum_{i=1}^m x_i \log x_i$. Then

$$B_F(p, q) = \sum_{i=1}^{m} \left(p_i \log \left(\frac{p_i}{q_i} \right) - p_i + q_i \right)$$

1.2. Properties

Proposition 1.3. The following statements are true for Bregman Divergences:

- Positive Definiteness: $B_F(p, q) \geq 0$ for all p, q, with equality if and only if $p = q$.
- Convexity: $B_F(p, q)$ is convex in its first argument.

- Conic Property: For $\alpha, \beta > 0$ we have $B_{\alpha F_1 + \beta F_2}(p, q) = \alpha B_{F_1}(p, q) + \beta B_{F_2}(p, q)$.
- Gradient in first argument: $\nabla_p B_F(p, q) = \nabla F(p) - \nabla F(q)$.

Proof. The Non-negativity of B_F is precisely the 1st order characterization of the convexity of F. The convexity of B_F in the first argument follows from the convexity of F. The Conic property follows immediately from the definition. If we allowed non-convex generators then we would not require the condition $\alpha, \beta > 0$, and this property would instead be linearity. Many authors refer to this property as linearity anyway, despite requiring strictly convex generators. The Gradient follows easily from the definition. □

Given a strictly convex function F, a naturally related strictly convex function is its convex conjugate (or Legendre-Fenchel transform). It turns out that the Bregman Divergences which they generate are closely related as well. First we recall a property of conjugates.

Lemma 1.4. Let $F^*(q) = \sup_{p \in \Delta}(\langle p, q \rangle - F(p))$ denote the convex conjugate of F which is strictly convex and continuously differentiable. Then for all $p \in \Delta$, we have

$$(\nabla F^*)(\nabla F(p)) = p$$

so ∇F is an invertible function from Δ onto its range, with its inverse being ∇F^*. Also, $\mathrm{dom}(F^*)$ is the range of ∇F.

Proof. Take a fixed $p \in \Delta$. By the first order condition, p is the maximizer in the definition $F^*(q) = \sup_{p \in \Delta}(\langle p, q \rangle - F(p))$, if and only if $q = \nabla F(p)$ (otherwise the supremum is infinite, so $\mathrm{dom}(F^*)$ is the range of ∇F). We have $F^*(q) + F(p) = \langle p, q \rangle$. Since F is continuous and strictly convex, we have $F = F^{**}$ by the Biconjugation theorem ([**3**]). Thus, we have

$$F^*(q) + F^{**}(p) = \langle p, q \rangle \tag{1.5}$$

implying that q is the maximizer in $F^{**}(p) = \sup_{r \in \Delta}(\langle p, r \rangle - F^*(r))$. The first order condition then implies that $p = \nabla F^*(q) = \nabla F^*(\nabla F(p))$. □

Theorem 1.6. Let F be strictly convex and continuously differentiable. Then for all $p \in \Delta, q \in \mathrm{dom}(F^*)$, we have

$$B_{F^*}(\nabla F(q), \nabla F(p)) = B_F(p, q).$$

Proof. By (1.5) we have $F^*(\nabla F(p)) = \langle p, \nabla F(p) \rangle - F(p)$, which lets us compute

$$\begin{aligned} B_{F^*}(\nabla F(q), \nabla F(p)) &= F^*(\nabla F(q)) - F^*(\nabla F(p)) - \langle \nabla F^*(\nabla F(p)), \nabla F(q) - \nabla F(p) \rangle \\ &= \langle q, \nabla F(q) \rangle - F(q) - \langle p, \nabla F(p) \rangle + F(p) - \langle p, \nabla F(q) - \nabla F(p) \rangle \\ &= F(p) - F(q) - \langle \nabla F(q), p - q \rangle \\ &= B_F(p, q). \end{aligned}$$

□

The results and examples seen above indicate that Bregman Divergences seem to capture some notion of distance. While they do not define metrics, they sometimes behave like a 'squared metric'. We now prove an important result that connects whatever notion of distance Bregman Divergences capture to means.

Theorem 1.7. (Mean is the Minimizer)
Suppose P is any random variable taking values in Δ with distribution $\mathcal{D}$. Then $q \mapsto \mathbb{E}_{P \sim \mathcal{D}}[B_F(P, q)]$ is minimized uniquely at $q^* = \mathbb{E}_{\mathcal{D}}[P]$. Note that this minimizer is independent of F.

Proof. Let q be any fixed point in Δ. Applying the definitions and linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}[B_F(P,q)] - \mathbb{E}[B_F(P,q^*)] &= \mathbb{E}[F(P) - F(q) - \langle \nabla F(q), P-q\rangle - F(P) + F(q^*) + \langle \nabla F(q^*), P - q^*\rangle] \\
&= F(q^*) - F(q) + \langle \nabla F(q), q\rangle - \langle \nabla F(q^*), q^*\rangle \\
&\quad + \mathbb{E}[\langle \nabla F(q^*), P\rangle - \langle \nabla F(q), P\rangle] \\
&= F(q^*) - F(q) + \langle \nabla F(q), q\rangle - \langle \nabla F(q^*), q^*\rangle \\
&\quad + \langle \nabla F(q^*) - \nabla F(q), \mathbb{E}[P]\rangle \\
&= F(q^*) - F(q) - \langle \nabla F(q), q^* - q\rangle \\
&= B_F(q^*, q)
\end{aligned}$$

Since $B_F(q^*, q) \geq 0$ with equality if and only if $q = q^*$, we have the result. □

The following converse result also holds.

Theorem 1.8. (Banerjee et al, 2005, [**2**])
Suppose $\Delta \subseteq \mathbb{R}^m$ is a closed convex set and $d : \Delta \times \Delta \to \mathbb{R}$ is a continuously differentiable positive definite function. If $q = \mathbb{E}[P]$ is the unique minimizer of $\mathbb{E}[d(P,q)]$ for all random variables P taking values in Δ, there exists a strictly convex differentiable function $F : \Delta \to \mathbb{R}$ such that $d = B_F$.

1.3. Optimization

We have seen that Bregman Divergences convey a notion of distance and that they are convex functions in their first argument. Thus, a natural optimization problem that often arises is to find the *Bregman projection* of a vector $q_0 \in \Delta \subseteq \mathbb{R}^m$ onto a linear subspace. Suppose we wish to find the $p \in \Delta$ which has least Bregman divergence to q_0 and satisfies some linear equality constraints. We explore a specific case below.

Problem– Suppose F is continuously differentiable. Further assume that the domain of F^* is $\mathbb{R}^m$, which by Lemma 1.4 is equivalent to $\nabla F : \Delta \to \mathbb{R}^m$ being a bijection. Let $\mathcal{P} = \{p \in \Delta \mid p^T M = \tilde{p}^T M\}$ be the set of points satisfying the constraints, where M is some $m \times n$ matrix and $\tilde{p} \in \Delta$. Note $\tilde{p} \in \mathcal{P}$ so $\mathcal{P}$ is non-empty. We wish to find $q_* = \arg\min_{p \in \mathcal{P}} B_F(p, q_0)$.

Towards solving the problem, let us define the *Legendre-Bregman projection* $\mathcal{L} : \Delta \times \mathbb{R}^m \to \Delta$ as

$$\mathcal{L}_F(q, v) = (\nabla F)^{-1}(\nabla F(q) - v),$$

which is well defined under our assumptions on F. Let $\mathcal{Q} := \{\mathcal{L}_F(q_0, M\lambda) \mid \lambda \in \mathbb{R}^n\} \subseteq \Delta$. Note $q_0 \in \mathcal{Q}$ so $\mathcal{Q}$ is non-empty.

The Lagrangian of our problem is given by

$$K(p, \lambda) = B_F(p, q_0) + (p^T M - \tilde{p}^T M)\lambda.$$

By the theory of Lagrange multipliers, the solution to the problem is given by the saddle point of the Lagrangian, where the minimum is taken with respect to p and the maximum with respect to λ. To minimize with respect to p we require $\nabla_p K(p, \lambda) = 0$ which we can write as

$$\nabla F(p) = \nabla F(q_0) - M\lambda.$$

Applying $(\nabla F)^{-1}$ to both sides yields $p = \mathcal{L}_F(q_0, M\lambda) \in \mathcal{Q}$. To maximize with respect to λ, we require $\nabla_\lambda K(p, \lambda) = 0$, which is simply the condition $p \in \mathcal{P}$. Therefore, solving the problem is equivalent to finding a point in $\mathcal{P} \cap \mathcal{Q}$.

Putting $\nabla F(p) = \nabla F(q_0) - M\lambda$ into the Lagrangian, the original problem is solved if we find the λ which maximizes $K(\mathcal{L}_F(q_0, M\lambda), \lambda)$, which can be simplified to

$$B_F(\tilde{p}, q_0) - B_F(\tilde{p}, \mathcal{L}_F(q_0, M\lambda)).$$

Note that the first term is a constant, and the second argument of the remaining term varies precisely over the points $q \in \mathcal{Q}$. So a dual problem to the original problem is to minimize $B_F(\tilde{p}, q)$ over $q \in \mathcal{Q}$, and both problems are solved if we find a point $q_* \in \mathcal{P} \cap \mathcal{Q}$. This opens the interesting possibility that we could prove some proposed $q_* \in \mathcal{P}$ is the solution of the original optimization problem simply by providing a λ which certifies $q_* \in \mathcal{Q}$.

An issue we have is that $\mathcal{P} \cap \mathcal{Q}$ may actually be empty and this approach does not yield a solution to our problem. It so happens that if we replace $\mathcal{Q}$ by its closure $\overline{\mathcal{Q}}$ (which means we replace the certificate λ with a sequence of λ_t, where the sequence itself may not converge to a finite value but $\mathcal{L}_F(q_0, M\lambda_t)$ does), then it can be shown that for a large class of F, the results we developed above still hold and $\mathcal{P} \cap \overline{\mathcal{Q}}$ contains exactly one point, the unique solution to the original and dual optimization problems.

Theorem 1.9. (Lafferty, Della Pietra, Della Pietra, 2001 [**5**])
Suppose $\Delta \subseteq \mathbb{R}^m$ is a compact convex set, and $F : \Delta \to \mathbb{R}$ is a strictly convex, continuously differentiable function whose convex conjugate has domain $\mathbb{R}^m$. Then there exists a unique $q_* \in \Delta$ satisfying:
1) $q_* \in \mathcal{P} \cap \overline{\mathcal{Q}}$.
2) $q_* = \arg\min_{q \in \overline{\mathcal{Q}}} B_F(\tilde{p}, q)$.
3) $q_* = \arg\min_{p \in \mathcal{P}} B_F(p, q_0)$.
Further, any of these properties determines q_* uniquely.

Remark 1.10. In the referenced paper, the conditions on F are somewhat more relaxed than we have stated here. The compactness condition on Δ is the main regard where we have been much more restrictive. We have done this for two reasons. First, it implies several of the technical conditions on F which appear in the referenced paper, which considerably simplifies the statement of the theorem and relieves us of checking those conditions. Second, although this excludes many F which may be of interest, it happens to include the main one which we will utilize later when we discuss Logistic Regression.

In the same paper, the authors also describe a method of 'Auxiliary functions' for solving problems of the types in the above theorem. Although this method is not completely general as it requires some ingenuity and/or luck to create an appropriate function, it is applicable to a variety of problems of this type. In the final chapter we see how the minimization of the log loss of logistic regression can be cast as a problem as in Theorem 1.9, and we then use the method of Auxiliary functions to prove that a proposed algorithm converges to the optimal parameters.

CHAPTER 2

Exponential Families

Exponential families are families of distributions which arise frequently in statistics. They have many connections to Bregman Divergences and will also be relevant in our discussion of Logistic Regression. Many commonly occurring families of distributions are exponential families, such as the normal, log-normal, exponential, gamma and bernoulli distributions. While the definition may appear to be an arbitrary generalization of some common occurring distributions, their naturality can be justified by some important properties which they possess.

Definition 2.1. A family of distributions parametrized by some vector θ is said to be an *exponential family* if their densities can be written in the form

$$p_A(x|\theta) = h(x)\exp(\langle F(\theta), t(x)\rangle - A(\theta))$$

If the densities can be written in this form, we call $t(x)$ a *sufficient statistic*, $F(\theta)$ the *natural parameters*, $A(\theta)$ the *log-normalizer* and $h(x)$ the *base measure* of the family.

Example 2.2. The normal family of distributions of with mean μ and variance σ^2 is an exponential family. We can verify that $h(x) = (2\pi)^{-1/2}, t(x) = (x, x^2), F(\mu, \sigma^2) = (\mu/\sigma^2, -1/(2\sigma^2))$, and $A(\mu, \sigma^2) = \mu^2/(2\sigma^2) + \log\sigma$ gives the desired distribution.

Example 2.3. The Poisson distribution with parameter λ arises as an exponential family by taking $h(x) = 1/x!, t(x) = x, F(\lambda) = \log\lambda$, and $A(\lambda) = \lambda$.

The most comprehensive source listing exponential families and their parametrizations, sufficient statistics etc, is Wikipedia. Many examples and an overview of their properties can also be found in Nielsen and Garcia [**9**].

2.1. Basic Properties

From now on, unless stated otherwise we assume that the parameters θ are already chosen to be the natural parameters, so

$$p_A(x|\theta) = h(x)\exp(\langle\theta, t(x)\rangle - A(\theta)).$$

Note that since p_A is a density, we have

$$A(\theta) = \log\left(\int_{\mathcal{X}} h(x)\exp(\langle\theta, t(x)\rangle)dx\right)$$

where $\int_{\mathcal{X}}$ is an integral over $\mathbb{R}^n$ for a continuous random variable and a summation for a discrete random variable. The *natural parameter* space Θ is the set of θ for which this expression is finite. The families with non-empty Θ is said to be *regular*. By directly verifying the definition we see that Θ is a convex set as well.

Differentiating this gives

$$\begin{aligned}\frac{\partial A(\theta)}{\partial \theta_i} &= \frac{\int_{\mathcal{X}} t(x)_i h(x) \exp(\langle \theta, t(x) \rangle) dx}{\int_{\mathcal{X}} h(x) \exp(\langle \theta, t(x) \rangle) dx} \\ &= \int_{\mathcal{X}} t(x)_i h(x) \exp(\langle \theta, t(x) \rangle - A(\theta)) dx \\ &= \mathbb{E}(t(x)_i)\end{aligned}$$

so we have $\mathbb{E}(t(x)) = \nabla A(\theta)$. Through similar steps we can compute

$$\frac{\partial^2 A(\theta)}{\partial \theta_i \partial \theta_j} = \operatorname{cov}(t(x)_i, t(x)_j),$$

so the Hessian of $A(\theta)$ is positive semi-definite at every point on its convex domain i.e. it is a convex function. Therefore it has a well defined convex conjugate

$$A^*(\eta) = \sup_{\theta \in \Theta} \left(\langle \theta, \eta \rangle - A(\theta) \right).$$

Differentiating shows that the supremum is attained when $\eta = \nabla A(\theta) = \mathbb{E}(t(x))$. Thus, the η are known as the *expectation parameters* or *moment parameters* of the exponential family. Since $A(\theta)$ is a continuous convex function, the Fenchel biconjugation theorem ([**3**]) gives that $A^{**} = A$. By the same reasoning as the previous step, we have $\theta = \nabla A^*(\eta)$. Therefore, to specify a member of a given exponential family, it suffices to give either the natural parameters or the expectation parameters, which can be calculated from one another. In other words, the mean of the sufficient statistic $\mathbb{E}(t(x))$ is in one to one correspondence with the natural parameter θ. Further, we have $\eta = \nabla A(\theta) = \nabla A(\nabla A^*(\eta))$ and $\theta = \nabla A^*(\eta) = \nabla A^*(\nabla A(\theta))$, i.e. $\nabla A(\theta)$ and $\nabla A^*(\eta)$ are inverse functions.

2.2. Inference

2.2.1. Maximum Likelihood Estimates and Method of Moments.

Suppose we have m i.i.d. samples $x_1, \ldots, x_m$ from a member of some exponential family

$$p_A(x|\theta) = h(x) \exp(\langle \theta, t(x) \rangle - A(\theta))$$

Then the likelihood satisfies

$$\begin{aligned}L(\theta) &= \prod_{i=1}^{m} h(x_i) \exp(\langle \theta, t(x_i) \rangle - A(\theta)) \\ &\propto \exp\left(\left\langle \theta, \sum_{i=1}^{m} t(x_i) \right\rangle - mA(\theta) \right)\end{aligned}$$

Setting the derivative of the log likelihood to zero, we see that the MLE parameter $\hat{\theta}$ satisfies

$$\nabla A(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^{m} t(x_i)$$

Since $\nabla A(\theta) = \mathbb{E}(t(x))$, we see that for exponential families Maximum Likelihood Estimation produces the same parameters as the method of matching moments of a sufficient statistic. As ∇A^* is inverse to ∇A, we have the following equation for the MLE parameter:

$$\hat{\theta} = \nabla A^* \left(\frac{1}{m} \sum_{i=1}^{m} t(x_i) \right)$$

Example 2.4. The family of normal distributions can be written as the exponential family with log-normalizer $A(\theta) = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$ and sufficient statistic $t(x) = (x, x^2)$. The natural parameters are related to the usual parameters by $\theta_1 = \mu/\sigma^2, \theta_2 = -1/(2\sigma^2)$.

We calculate the gradient

$$\nabla A(\theta) = \left(\frac{-\theta_1}{2\theta_2},\ \frac{-1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} \right) = \eta$$

Solving for η_1, η_2 in terms of θ_1, θ_2, we have

$$\nabla A^*(\eta) = \theta = \left(\frac{-\eta_1}{\eta_1^2 - \eta_2},\ \frac{1}{2(\eta_1^2 - \eta_2)} \right)$$

Taking $\eta = \frac{1}{m}\sum_{i=1}^{m} t(x_i)$ gives

$$\hat{\theta} = \left(\frac{\hat{\mu}}{\hat{\sigma^2}},\ \frac{-1}{2\hat{\sigma^2}} \right)$$

where $\hat{\mu}, \hat{\sigma^2}$ are the sample mean and sample variance respectively. Thus, the MLE μ, σ^2 parameters are the sample mean and sample variance, as expected.

2.2.2. Bregman Distance between members of an exponential family.
Let $KL(p, q)$ denote the Kullback-Liebler divergence between two densities p and q. The KL divergence between two members of the same exponential family takes on a particularly elegant form.

$$\begin{aligned}
KL(p_A(x|\theta), p_A(x|\theta')) &= \int_{\mathcal{X}} p_A(x|\theta) \log \left(\frac{p_A(x|\theta)}{p_A(x|\theta')} \right) dx \\
&= \int_{\mathcal{X}} p_A(x|\theta) \log \left(\frac{\exp(\langle \theta, t(x) \rangle - A(\theta))}{\exp(\langle \theta', t(x) \rangle - A(\theta'))} \right) dx \\
&= \int_{\mathcal{X}} p_A(x|\theta) \left(\langle \theta - \theta', t(x) \rangle - A(\theta) + A(\theta') \right) dx \\
&= A(\theta') - A(\theta) - \int_{\mathcal{X}} p_A(x|\theta) \langle \theta' - \theta, t(x) \rangle dx \\
&= A(\theta') - A(\theta) - \left\langle \theta' - \theta, \int_{\mathcal{X}} p_A(x|\theta) t(x) dx \right\rangle \\
&= A(\theta') - A(\theta) - \langle \theta' - \theta, \mathbb{E}(t(x)) \rangle \\
&= A(\theta') - A(\theta) - \langle \theta' - \theta, \nabla A(\theta) \rangle \\
&= B_A(\theta', \theta)
\end{aligned}$$

So the KL divergence between two members of the same exponential family can be expressed as a Bregman Divergence generated by their log-normalizer evaluated at their natural parameters.

2.3. Natural Exponential Families

We now define a special class of exponential families which we will refer to in the next section.

Definition 2.5. We say an exponential family is a *natural exponential family* if $t(x) = x$ is a sufficient statistic for the distributions. Writing the densities in terms of natural parameters, members of a natural exponential family have densities of the form

$$p_A(x|\theta) = h(x) \exp(\langle x, \theta \rangle - A(\theta))$$

Example 2.6. Important examples of normal exponential families include the normal distribution with known covariance, gamma distribution with known shape parameter, binomial distribution with known number of trials, the Poisson distribution, and the negative binomial distribution with known r. These include other distributions as special cases, such as the exponential, Bernoulli, geometric and chi-squared distributions. Important non-examples are the beta distribution and log-normal distribution (which are exponential families, but not natural exponential families).

Some of the results developed in the previous section simplify further. For natural exponential families, we have $\mathbb{E}(x) = \nabla A(\theta)$ and $\operatorname{cov}(x) = \nabla^2 A(\theta)$. The maximum likelihood estimate for an i.i.d. sample is given by $\hat{\theta} = \nabla A^* (\hat{x})$, where $\hat{x}$ is the sample mean. See [**8**] for more advanced results.

CHAPTER 3

Logistic Regression

In this chapter we discuss the problems of Binary Classification and Regression, noting their similarities. Extending the ideas of Linear Regression, we define Generalized Linear Models and show how the Logistic Regression model naturally arises.

3.1. Binary Classification

Suppose $\mathcal{D}$ is a distribution on $\mathbb{R}^n \times \{\pm 1\}$ with a density $p(x, y)$. Binary classification is the task of determining the class probability function $\mathbb{R}^n \to [0, 1] : x \mapsto p(y = +1|x)$.

Example 3.1. Suppose that $\mathcal{D}$ is a mixture of two normal distributions. More precisely, assume that $p(x|y = +1), p(x|y = -1)$ have the same covariance matrix Σ and means μ^+ and μ^- respectively, and that the marginals $p(y = +1), p(y = -1)$ are known[1]. We have the algebraic identity

$$p(y = +1|x) = \frac{1}{1 + \exp\left(-\log\left(\frac{p(y=+1|x)}{p(y=-1|x)}\right)\right)} \tag{3.2}$$

$$= \sigma\left(\log\left(\frac{p(y = +1|x)}{p(y = -1|x)}\right)\right) \tag{3.3}$$

where $\sigma : \mathbb{R} \longrightarrow [0, 1] : x \mapsto \frac{1}{1+\exp(-x)}$ is the logistic function. [2] In this particular example we can explicitly calculate the log odds-ratio:

$$\begin{aligned}
\log\left(\frac{p(y = +1|x)}{p(y = -1|x)}\right) &= \log\left(\frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)}\right) \\
&= \log\det(2\pi\Sigma) - \frac{1}{2}(x - \mu^+)^T\Sigma^{-1}(x - \mu^+) \\
&\quad - \log\det(2\pi\Sigma) + \frac{1}{2}(x - \mu^+)^T\Sigma^{-1}(x - \mu^+) + \log\frac{p(y = +1)}{p(y = -1)} \\
&= x^T\Sigma^{-1}(\mu^+ - \mu^-) - \frac{1}{2}\left((\mu^+)^T\Sigma^{-1}\mu^+ - (\mu^-)^T\Sigma^{-1}\mu^-\right) + \log\frac{p(y = +1)}{p(y = -1)}
\end{aligned}$$

Note that this is an affine function of x, and combining this with the above identity allows us to easily calculate $p(y = +1|x)$, which solves the binary classification problem.

Remark 3.4. In Machine Learning and Statistics, we rarely have such an explicit description of $\mathcal{D}$. Instead, the information we have about $\mathcal{D}$ is usually in the form of a sequence of samples from the

[1] Note that these are precisely the assumptions of Linear Discriminant Analysis, but we proceed differently.

[2] For the next section it is helpful to note that this also gives us an expression for $\mathbb{E}(y|x)$ via the equation $\mathbb{E}(y|x) = 2p(y = +1|x) - 1$.

distribution. We can hardly ever assume the distribution is the mixture of two normal distributions with equal covariance matrix either. Nevertheless, the following sections will show the usefulness of this example.

3.2. Generalized Linear Models

Suppose $\mathcal{D}$ is a distribution on $\mathbb{R}^n \times \mathbb{R}$ with a density $p(x, y)$. Regression is the task of determining the function $\mathbb{R}^n \to \mathbb{R} : x \mapsto p(y|x)$.

In Linear Regression, we make the assumption that the distributions $p(y|x)$ are normal distributions of various means and a fixed variance. We then assume that $\mathbb{E}(y|x) = \langle \lambda, x \rangle + \lambda_0$ (i.e. an affine function of x). This model can be adjusted so that it becomes applicable to problems other than Regression.

Definition 3.5. A *Generalized Linear Model* (GLM) is determined by two components. First we specify an exponential family which we assume the densities $p(y|x)$ are members of. Then we specify an invertible function $g : S \longrightarrow \mathbb{R}$ whose domain $S \subseteq \mathbb{R}$ contains all possible values of $\mathbb{E}(y|x)$. The inverse g^{-1} is called the *link function*. The GLM is then the model

$$\mathbb{E}(y|x) = g^{-1}(\langle \lambda, x \rangle + \lambda_0)$$

Remark 3.6. In the cases where g^{-1} is smooth and $p(y|x)$ is determined by the mean $\mathbb{E}(y|x)$ (such as in Linear Regression and Example 3.1), the likelihood of a sample of i.i.d. data is a smooth function of the parameters. We can explicitly calculate derivatives in order to apply optimization methods such as Gradient Descent or Newton's method to calculate the MLE parameters[3]. When creating GLMs we attempt to find a link function such that this optimization problem is convex.

Writing $g\left(\mathbb{E}(y|x)\right) = \langle \lambda, x \rangle + \lambda_0$ makes the fundamental idea behind GLMs more evident. In the Regression problem, where $\mathbb{E}(y|x)$ may range over all values in $\mathbb{R}$, it can be possible to model it with an affine function of x. A barrier to applying the ideas of Linear Regression to other problems is that approximating $\mathbb{E}(y|x)$ by an affine function may be inappropriate. For example, in the Binary Classification problem, $\mathbb{E}(y|x)$ is bounded in $[-1, 1]$ whereas the only affine functions which are bounded are constant. Thus, we try to find a g which maps $\mathrm{domain}(\mathbb{E}(y|x))$ surjectively onto $\mathbb{R}$ such that this composition is well approximated by an affine function.

The set of appropriate functions g is clearly dependent on the distributions $p(y|x)$. Much of the introductory study of GLMs focuses on finding appropriate link functions in the special cases where the distributions $p(y|x)$ are assumed to be members of a certain exponential family (e.g. see **[6]** for many examples and general theory on GLMs).

3.3. Logistic Regression

By viewing Example 3.1 in the GLM framework, the following definition naturally arises.

[3]This will obtain the parameters which in some sense best explain the previously observed data. In practice, we are often more interested in making predictions on unobserved data. The MLE parameters will be finely fit to the observed data, and if the the observed data does not represent a sufficiently representative sample of the true distribution $\mathcal{D}$, then we may find that the MLE parameters are 'overfit' to the observed data and give inaccurate predictions on new data. To address this, we often try to extend GLMs to include regularization or Bayesian methods.

Definition 3.7. *Logistic Regression* is the GLM where the densities $p(y|x)$ are from the Bernoulli family and with the link function $g^{-1} = 2\sigma - 1$. In this case, the model can be written as

$$p(y = +1|x) = \sigma(f_\lambda(x))$$

where $f_\lambda(x) = \lambda_0 + \sum_{j=1}^{n} \lambda_j x_j$.

Given i.i.d. samples $(x_i, y_i), i = 1, \ldots, m$, the likelihood of the observed sample is given by

$$\prod_{i=1}^{m} \frac{1}{1 + \exp(-y_i f_\lambda(x_i))}$$

Maximizing this is equivalent to minimizing the *log loss*

$$\sum_{i=1}^{m} \log(1 + \exp(-y_i f_\lambda(x_i))$$

Note that $\log(1+\exp(x))$ is a convex function since its second derivative is $\sigma(x)(1-\sigma(x)) \geq 0$. The $-y_i f_\lambda(x_i)$ terms are affine in λ, so each summand is convex in λ. Thus, minimizing the log loss is a convex optimization problem, and if it attains its global minimum at some point $\lambda^* \in \mathbb{R}^n$ then methods such as Newton's method are guaranteed to converge to that point. However, the log loss may not attain a global minimum. This would be the case if the data is strictly linearly separable, i.e. if there exists λ such that $y_i(f_\lambda(x_i)) > 0$ for all $i = 1, \ldots, m$. In this situation, repeatedly scaling λ by some factor $\alpha > 1$ would bring each term in the log loss arbitrarily close to 0, while the non-zero components of λ would diverge (although the decision boundary $f_\lambda(x) = 0$ would remain unchanged).

In Example 3.1, where we assumed that the $p(x|y = +1), p(x|y = -1)$ are normal distributions with the same covariance matrix, we had

$$p(y = +1|x) = \sigma(\langle \lambda, x \rangle + \lambda_0)$$

for some parameters λ_0, λ, so in that case Logistic Regression is an 'accurate' GLM (in the sense that $p(y = +1|x)$, and hence $g(\mathbb{E}(y|x))$, is indeed an affine function of x). We now give a characterization of the distributions $p(x|y = +1), p(x|y = -1)$ for which the Logistic Regression model is 'accurate'.

Theorem 3.8. (A. Banerjee, [**1**])
The log-odds ratio of the class posteriors is affine if and only if the class conditional distributions belong to a fixed natural exponential family.

Proof. First suppose that the class conditional distributions $p(x|y = +1), p(x|y = -1)$ belong to the fixed natural exponential family with log-partition function $A(\theta)$ and base measure $h(x)$, with natural parameters θ_+, θ_- respectively. Then we have

$$\begin{aligned}\log\left(\frac{p(y=+1|x)}{p(y=-1|x)}\right) &= \log\left(\frac{p(x|y=+1)p(y=+1)}{p(x|y=-1)p(y=-1)}\right)\\ &= \log p(x|y=+1) - \log p(x|y=-1) + \log\frac{p(y=+1)}{p(y=-1)}\\ &= \langle\theta_+ - \theta_-, x\rangle - A(\theta_+) + A(\theta_-) + \log\frac{p(y=+1)}{p(y=-1)}\end{aligned}$$

which is indeed affine in x. Conversely, suppose the log-odds ratio is affine, i.e.

$$\log\left(\frac{p(y=+1|x)}{p(y=-1|x)}\right) = \log\left(\frac{p(x|y=+1)p(y=+1)}{p(x|y=-1)p(y=-1)}\right) = \langle a, x\rangle + b$$

for some constants a, b. Exponentiating both sides of this equation gives

$$p(x|y=+1) = p(x|y=-1)\exp(\langle a, x\rangle + b - \alpha)$$

where $\alpha := \log\frac{p(y=+1)}{p(y=-1)}$. As the integral of the right hand side over all x must be equal to 1, we get the equality

$$b - \alpha = -\log\left(\int_{\mathbb{R}^n}\exp(\langle a, x\rangle)p(x|y=-1)dx\right).$$

Therefore, if we define $A(\theta) = \log\left(\int_{\mathbb{R}^n}\exp(\langle\theta, x\rangle)p(x|y=-1)dx\right)$, then we have

$$p(x|y=+1) = p(x|y=-1)\exp(\langle a, x\rangle - A(a))$$

and

$$p(x|y=-1) = p(x|y=-1)\exp(\langle 0, x\rangle - A(0)).$$

That is, the class conditional distributions $p(x|y+=1), p(x|y=-1)$ are members of the natural exponential family characterized by base measure $p(x|y=-1)$ and log-partition $A(\theta)$, with natural parameters a and 0 respectively.

□

Remark 3.9. This theorem is due to A. Banerjee ([**1**]). In his proof he invoked some facts about Laplace transforms, and in the converse direction he split the proof into two cases with various assumptions on $p(x|y=-1)$, both of which we found inessential to the core argument so we have given a simpler proof above. Also note that the definition of 'Exponential Families' appearing in that paper refers to what is now more commonly known as 'Natural Exponential Families'. Due to this, his statement of this theorem does not include the term 'Natural', which could lead an unaware reader to think that Logistic Regression is accurate in a greater class than is true.

Remark 3.10. This theorem allows us to view Logistic Regression as a more 'robust' model than Linear Discriminant Analysis in the sense that the assumptions of Logistic Regression are true in a strict superset of the cases where the assumptions of Linear Discriminant Analysis are true.

CHAPTER 4

Logistic Regression via Bregman Distance Optimization

In this chapter, we given an exposition of results by Collins, Schapire and Singer ([**4**]), which puts the optimization of the log loss of the Logistic Regression model into the framework of a Bregman Divergence optimization problem of the type discussed in Chapter 1. This then allows them to present a new algorithm to determine the minimizing parameters of the log loss and elegantly prove the convergence of that algorithm.

4.1. Preliminaries

Suppose $\mathcal{D}$ is a distribution on $\mathbb{R}^n \times \{\pm 1\}$, and $(x_i, y_i), i = 1, \ldots, m$ are i.i.d samples from that distribution. We return to the setting and notation of Theorem 1.9.

Let $\Delta = [0,1]^m$. Let $\tilde{p} = \mathbf{0}, q_0 = (1/2)\mathbf{1}$, and $M_{ij} = y_i x_{ij}$ (where x_{ij} is the j-th component of the i-th sample). Let

$$F(p) = \sum_{i=1}^{m} \left(p_i \log p_i + (1 - p_i) \log(1 - p_i)\right).$$

The partial derivative with respect to p_i is equal to $\log(p_i(1-p_i)^{-1}) = \sigma^{-1}(p_i)$, so the gradient $\nabla F(p)$ is surjective onto $\mathbb{R}^m$ and the conditions of Theorem 1.9 are satisfied. With our calculation of the gradient, we have

$$\begin{aligned} \mathcal{L}_F(q, v)_i &= [(\nabla F)^{-1}(\nabla F(q) - v)]_i \\ &= \frac{q_i e^{-v_i}}{1 - q_i + q_i e^{-v_i}} \end{aligned}$$

so that

$$\mathcal{Q} = \left\{ q \in [0,1]^m \;\middle|\; q_i = \sigma\left(-\sum_{j=1}^{n} \lambda_j y_i x_{ij}\right), \lambda \in \mathbb{R}^n \right\}.$$

We also have

$$B_F(p, q) = \sum_{i=1}^{m} \left(p_i \log\left(\frac{p_i}{q_i}\right) + (1 - p_i) \log\left(\frac{1 - p_i}{1 - q_i}\right)\right)$$

so $B_F(\mathbf{0}, q) = -\sum_{i=1}^{m} \log(1 - q_i)$. Thus,

$$B_F(\mathbf{0}, \mathcal{L}_F(q_0, M\lambda)) = \sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \sum_{j=1}^{n} \lambda_j x_{ij}\right)\right)$$

so minimizing log-loss is equivalent to minimizing $B_F(\mathbf{0}, q)$ over $q \in \overline{\mathcal{Q}}$, which is a problem of type 2 in Theorem 1.9.

4.2. Algorithm

Parameters: $\Delta = [0,1]^m$, $q_0 = (1/2)\mathbf{1}$, and $F(p) = \sum_{i=1}^m \left(p_i \log p_i + (1-p_i)\log(1-p_i)\right)$.

Input: $M_{ij} = y_i x_{ij}$ where WLOG the x components of the sample data has been scaled so that $\sum_{j=1}^n |M_{ij}| \leq 1$ for all i, and the data is *regular* such that $W_{t,j}^+, W_{t,j}^-$ calculated below are never equal to 0.

Output: $(\lambda_t)_{t=1,2,\ldots}$ such that

$$\lim_{t\to\infty} B_F(\mathbf{0}, \mathcal{L}_F(q_0, M\lambda_t)) = \inf_{\lambda\in\mathbb{R}^n} B_F(\mathbf{0}, \mathcal{L}_F(q_0, M\lambda)).$$

Algorithm:
Let $\lambda_1 = \mathbf{0}$.
For $t = 1, 2, 3, \ldots$:
- $q_t = \mathcal{L}_F(q_0, M\lambda_t)$
- For $j = 1, 2, \ldots, n$:

$$W_{t,j}^+ = \sum_{i:\text{sign}(M_{ij})=+1} q_{t,i}|M_{ij}|$$

$$W_{t,j}^- = \sum_{i:\text{sign}(M_{ij})=-1} q_{t,i}|M_{ij}|$$

$$\delta_{t,j} = \frac{1}{2}\log\left(\frac{W_{t,j}^+}{W_{t,j}^-}\right)$$

- Let $\lambda_{t+1} = \lambda_t + \delta_t$.

Remark 4.1. There is a simple characterization of when the data is regular in the sense defined above. We can then extend Collins, Schapire and Singer's algorithm to handle cases where the data is not regular as well. By induction, $q_{t,i} > 0$. So $W_{t,j}^+ = 0$ iff $\sum_{i:y_i x_{ij}\geq 0} |x_{ij}| = 0$ iff ($y_i x_{ij} \geq 0$ implies $x_{ij} = 0$.)
Similarly, $W_{t,j}^- = 0$ iff $\sum_{i:y_i x_{ij}<0} |x_{ij}| = 0$ iff there are no i such that $y_i x_{ij} < 0$. These fact imply

- If $W_{t,j}^+ = 0$ and $W_{t,j}^- = 0$:
  Then we have $x_{ij} = 0$ for all i. In this case, the log-loss on the observed data does not depend on the value of λ_j, and we fix $\lambda_j = 0$.
- If $W_{t,j}^- = 0$ and $W_{t,j}^+ > 0$:
  Then we have $y_i x_{ij} \geq 0$ for every sample number $i = 1, \ldots, n$, and $x_{ij} \neq 0$ for at least one value of i. We can always reduce the log-loss on the observed data

$$\sum_{i=1}^m \log\left(1 + \exp\left(\sum_{j=1}^n (-\lambda_j)(y_i x_{ij})\right)\right)$$

  by increasing λ_j, so we fix $\lambda_j = +\infty$.
- If $W_{t,j}^+ = 0$ and $W_{t,j}^- > 0$: There is at least one i such that $y_i x_{ij} < 0$, and for any i where that is not the case we have $y_i x_{ij} = 0$. We can always reduce the log-loss on the observed data

$$\sum_{i=1}^{m} \log\left(1 + \exp\left(\sum_{j=1}^{n} \lambda_j(-y_i x_{ij})\right)\right)$$

by decreasing λ_j, so we fix $\lambda_j = -\infty$.

Then for each j which causes the regular to not be regular we strip the j-th column from the data matrix and apply the above algorithm to calculate the remaining parameters.

4.3. Proof of Convergence

Definition 4.2. For the sequence (q_t) and matrix M above, define an *auxiliary function* to be a continuous function $A : \Delta \to \mathbb{R}$ which satisfies the conditions that

$$B_F(0, q_{t+1}) - B_F(0, q_t) \leq A(q_t) \leq 0, \quad \forall t = 1, 2, \ldots$$

and

$$A(q) = 0 \implies q^T M = 0.$$

Lemma 4.3. Let A be an auxiliary function for (q_t) and M as in the above algorithm. Then

$$\lim_{t\to\infty} q_t = q_* := \arg\min_{q\in\overline{\mathcal{Q}}} B_F(0, q).$$

Thus, proving the convergence of the algorithm is reduced to finding an auxiliary function for the sequence (q_t) and matrix M above.

Proof. From the first condition of an auxiliary function, the sequence $B_F(0, q_{t+1})$ is non-increasing. It is also bounded from below by 0. Every non-increasing sequence bounded from below converges to a finite limit. Now the squeeze theorem applied to the first condition of auxiliary functions implies that $A(q_t)$ converges to 0. Recall the Bolzano-Weierstrass theorem, that every bounded sequence in $\mathbb{R}^m$ has a convergent subsequence. Since Δ, is compact, (q_t) has a convergent subsequence converging to some limit point $\hat{q} \in \overline{\mathcal{Q}}$. Continuity of A then implies that $A(\hat{q}) = 0$. The second condition of auxiliary functions then implies that $\hat{q}$ is such that $\hat{q}^T M = 0 = \tilde{p}^T M$, so $q \in \mathcal{P}$. Since $\hat{q} \in \mathcal{P} \cap \overline{\mathcal{Q}}$, by Theorem 1.9 we have $\hat{q} = q_*$. This reasoning applies for any limit point $\hat{q}$, and Theorem 1.9 states that q_* is unique, so the sequence (q_t) has a unique limit point. Given this, suppose (q_t) does not converge to q_*, meaning that these exists an open set U around q_* such that infinitely many points of (q_t) are not in U. Then these points have some limit point in $\Delta \setminus U$ (which is closed). This contradicts (q_t) having a unique limit point, so the sequence (q_t) converges to q_*. □

Theorem 4.4. For any $q \in \Delta$, let

$$W_j^+(q) = \sum_{i:\text{sign}(M_{ij})=+1} q_i |M_{ij}|$$

$$W_j^-(q) = \sum_{i:\text{sign}(M_{ij})=-1} q_i |M_{ij}|$$

The function

$$A(q) = -\sum_{j=1}^{n} \left(\sqrt{W_j^+(q)} - \sqrt{W_j^-(q)}\right)^2$$

is an auxiliary function for (q_t) and M as in the above algorithm, and therefore the algorithm converges.

Proof. First note that A is continuous and bounded from above by 0. Simple manipulation verifies that Legendre-Bregman projections satisfy the property $\mathcal{L}_F(\mathcal{L}_F(q,w),v) = \mathcal{L}_F(q,v+w)$. This gives

$$\begin{aligned} q_{t+1} &= \mathcal{L}_F(q_0, M(\lambda_t + \delta_t)) \\ &= \mathcal{L}_F(\mathcal{L}_F(q_0, M\lambda_t), M\delta_t) \\ &= \mathcal{L}_F(q_t, M\delta_t). \end{aligned}$$

We previously computed that for F as in the algorithm, $B_F(0,q) = -\sum_{i=1}^m \log(1-q_i)$. From this we have

$$\begin{aligned} B_F(0, \mathcal{L}_F(q,v)) - B_F(0,q) &= \sum_{i=1}^m \log\left(\frac{1-q_i}{1-\mathcal{L}_F(q,v)_i}\right) \\ &= \sum_{i=1}^m \log(1 - q_i + q_i e^{-v_i}) \\ &\leq \sum_{i=1}^m (-q_i + q_i e^{-v_i}) \end{aligned}$$

where the inequality followed from the fact that $e^x \geq 1 + x$ for all x. Therefore we have

$$\begin{aligned} B_F(0, q_{t+1}) - B_F(0, q_t) &= B_F(0, \mathcal{L}_F(q_t, M\delta_t)) - B_F(0, q_t) \\ &\leq \sum_{i=1}^m q_{t,i} \left(\exp\left(-\sum_{j=1}^n \delta_{t,j} s_{ij} |M_{ij}| \right) - 1 \right) \end{aligned}$$

where s_{ij} is the sign of M_{ij}. Note that for any x_j and for $p_j \geq 0$ with $\sum_{j=1}^n p_j \leq 1$, Jensen's inequality applied to $e^x - 1$ gives $\exp\left(\sum_{j=1}^n p_j x_j\right) - 1 \leq \sum_{j=1}^n p_j (e^{x_j} - 1)$. Therefore we have

$$B_F(0, q_{t+1}) - B_F(0, q_t) \leq \sum_{i=1}^m q_{t,i} \left(\sum_{j=1}^n |M_{ij}| (\exp(-\delta_{t,j} s_{ij}) - 1) \right) \tag{4.5}$$

$$= \sum_{j=1}^n \left(W_{t,j}^+ e^{-\delta_{t,j}} + W_{t,j}^- e^{\delta_{t,j}} - W_{t,j}^+ - W_{t,j}^- \right) \tag{4.6}$$

$$= -\sum_{j=1}^n \left(\sqrt{W_{t,j}^+} - \sqrt{W_{t,j}^-} \right)^2 \tag{4.7}$$

$$= A(q_t) \tag{4.8}$$

These steps followed by the definitions of $W_{t,j}^+, W_{t,j}^-$ and $\delta_{t,j}$ (which was specifically chosen to minimize (4.6)). We have verified that A satisfies the first condition required of an auxiliary function. Now suppose that $A(q) = 0$. Then we have $W_j^+(q) = W_j^-(q)$ for all j. Therefore

$$\begin{aligned} 0 &= W_j^+(q) - W_j^-(q) \\ &= \sum_{i=1}^{m} q_i s_{ij} |M_{ij}| \\ &= \sum_{i=1}^{m} q_i M_{ij} \\ &= (q^T M)_j \end{aligned}$$

which confirms the second condition as well. So A is an auxiliary function for (q_t) and M, proving the result.

□

Remark 4.9. We mentioned that the δ_t defined in Collins, Schapire and Singer's algorithm is chosen specifically to minimize equation (4.6). Suppose that instead we define δ_t to be the value which minimizes equation (4.6) subject to the constraint that $\|\lambda_t + \delta_t\|_1 \leq \alpha$ for some fixed parameter $\alpha > 0$. Finding δ_t is then a convex optimization problem to which we can apply well established methods to solve. This altered algorithm outputs a sequence of parameters λ_t for which $\|\lambda_t\|_1 \leq \alpha$ for every t, and if this sequence converges to some λ_* then we have $\|\lambda_*\|_1 \leq \alpha$ as well. In **[7]**, Huang and Gupta show that this altered algorithm converges to a λ which is a solution to the convex optimization problem of minimizing the log loss subject to the constraint that $\|\lambda\|_1 \leq \alpha$. This constraint on λ is known as Ivanov regularization.

Appendix

Python implementations of Collins, Schapire and Singer's algorithm for Logistic Regression, as well as Huang and Gupta's version with Ivanov regularization, are publically available at: *https://github.com/chengsoonong/eheye/blob/master/BregmanLR/notebook/*

We compared these algorithms against the more well known approaches of minimizing log loss by L-BFGS, and minimizing log loss under L1 Ivanov regularization by the Lasso method. Testing on a sample of well known datasets gave the following accuracy results.

Dataset	Instances	Features	LR	Bregman LR	LR+L1	Bregman LR+L1
MNIST	14780	784	99.8	99.6	99.1	99.9
Fashion MNIST	14000	784	98.4	98.8	98.9	97.0
Ionosphere	351	35	98.0	88.2	90.2	86.3
Diabetes	768	8	82.7	73.7	83.6	72.7
Heart Statlog	270	13	89.7	87.2	87.2	89.7
WDBC	569	30	98.8	90.2	95.1	86.6

References

[1] A. BANERJEE 'An Analysis of Logistic Models: Exponential Family Connections and Online Performance'. *2007 SIAM International Conference on Data Mining* 2007.

[2] A. BANERJEE, X. GOU, H. WANG 'On the Optimality of Conditional Expectation as a Bregman Predictor' *IEEE Trans. on Information Theory* Vol 51(7) 2005.

[3] J. BORWEIN AND A. LEWIS, *Convex Analysis and Nonlinear Optimization*, (2 ed.). Springer. p. 76, 2006.

[4] M. COLLINS, R.E. SCHAPIRE, Y. SINGER 'Logistic Regression, AdaBoost and Bregman Distances' *Machine Learning*, 48(1/2/3). 2002.

[5] S. DELLA PIETRA, V. DELLA PIETRA, J. LAFFERTY, 'Duality and Auxiliary Functions for Bregman Distances', *Tech. rep. CMU-CS-01-109, School of Computer Science, Carnegie Mellon University.* 2001.

[6] A. J. DOBSON AND A. G. BARNETT, *Introduction to Generalized Linear Models*, Boca Raton, FL: Chapman and Hall/CRC, 2008.

[7] T. HUANG, M. GUPTA Bregman distance to L1 regularized logistic regression. *International Conference on Pattern Recognition* 2008.

[8] C. MORRIS, "Natural Exponential Families with Quadratic Variance Functions", Annals of Statistics, 10 (1), p65-80. 1982

[9] F. NIELSEN, V. GARCIA, "Statistical exponential families: A digest with flash cards", arXiv:0911.4863v2, 2011.

[10] R. NOCK, A. MENON AND C. ONG, "A scaled Bregman theorem with applications", NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems, p. 19-27, 2016.