Local Likelihood Modeling of Temporal Text Streams

Guy Lebanon LEBANON@STAT.PURDUE.EDU
Yang Zhao ZHAO18@STAT.PURDUE.EDU
Department of Statistics, Purdue University – West Lafayette, IN 47907

Abstract

Temporal text data is often generated by a time-changing process or distribution. Such a drift in the underlying distribution cannot be captured by stationary likelihood techniques. We consider the application of local likelihood methods to generative and conditional modeling of temporal document sequences. We examine the asymptotic bias and variance and present an experimental study using the RCV1 dataset containing a temporal sequence of Reuters news stories.

1. Introduction

Time stamped documents such as news stories often cannot be accurately modeled by a single time invariant distribution. An alternative is to assume that the concepts underlying the distribution generating the data drift with time. In other words, the data is generated by a time dependent process $z^{(t)} \sim p_t(z), t \in I \subset \mathbb{R}$ whose approximation $\{\hat{p}_t : t \in I\}$ becomes the main objective of the learning task. We assume that the time t is a continuous quantity, even in cases where the realized time points form a discrete sample. For example, assuming that the time stamps represent the days of the year when the documents were authored, we assume that the set $\{1, \ldots, 365\}$ is a discrete sample from a underlying continuous interval $[1, 365]$. We further assume that the data samples $z^{(t)}$, sampled from p_t, correspond to pairs $z^{(t)} = (x, y)$ constituting a document x and a categorial-valued label y. Such pairs (x, y) appear often in practice, for example with y corresponding to the document topic (Lewis et al., 2004), sentiment (Pang & Lee, 2005), author (Mosteller & Wallace, 1964) or Email spam/no-spam (Mulligan, 1999).

Assuming that our data is a set of time stamped documents and labels $(t, (x, y))$, the drift $p_t(x, y)$ can be characterized by considering the temporal transition of the joint distribution $p_t(x, y)$, the conditionals $p_t(y|x)$, $p_t(x|y)$, or the marginals $p_t(x), p_t(y)$. The choice of which of the distributions above to model depends on the application at hand. For example, modeling $p_t(y|x)$ is usually sufficient for document classification purposes while modeling $p_t(x|y)$ is necessary for language modeling which is an important component in speech recognition, machine translation, and IR.

We demonstrate the presence of concept drift in practice by considering the Reuters RCV1 dataset (Lewis et al., 2004) which contains over 800,000 news stories gathered in a period spanning 365 consecutive days and categorized according to topic. Figure 1 displays the temporal change in the relative frequency (number of appearance in a document divided by document length) of three words: `million`, `common`, and `Handelsgesellschaft` (German trade unions) for documents in the most popular RCV1 category titled `CCAT`. It is obvious from these plots that the relative frequency of these words vary substantially in time. For example, the word `Handelsgesellschaft` appear in 8 distinct time regions, representing time points in which German trade unions were featured in the Reuters news archive.

The temporal variation in relative frequencies illustrated by Figure 1 corresponds to a drift in the distribution generating the data. Since the drift is rather pronounced, standard estimation methods based on maximum likelihood are not likely to accurately model the data. In this paper, we consider instead estimating $\{p_t(x, y) : t \in I\}$ based on the the local likelihood principle. Local likelihood is a locally weighted version of the loglikelihood with the weights determined by the difference between the time points associated with the sampled data and a the time at which the inference takes place.

After presenting a more formal discussion of concept drift in Section 3 and the definition of local likelihood in Section 4 we turn to examine in detail the case of

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

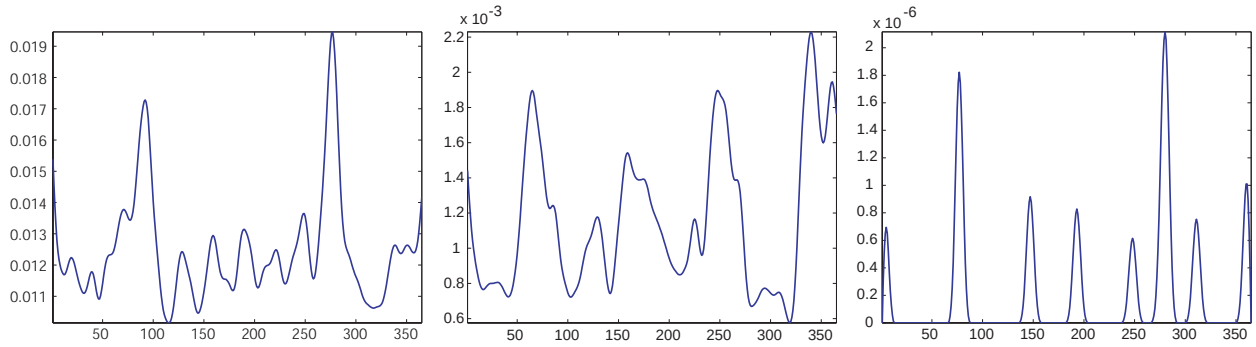

Figure 1. Estimated relative frequency (number of appearances in a document divided by document length) of three words from the most popular category in RCV1 as a function of time. The three panels correspond to the words `million`, `common`, and `Handelsgesellschaft` (German trade unions). The displayed curves were smoothed to remove sampling noise.

modeling $p_t(x|y)$ with local likelihood for n-grams and modeling $p_t(y|x)$ with local likelihood for logistic regression. In the case of 1-grams or the naive Bayes model, we provide a precise as well as asymptotic description of the bias and variance which illuminates certain facts concerning the selection of weights and the difference between the online and offline scenarios. Experiments conducted on the RCV1 dataset demonstrates the local likelihood estimation in practice and contrasts it with more standard non-local alternatives.

2. Related Work

Concept drift or similar phenomena under different names have been studied in a number of communities. It has recently gained interest primarily due to an increase in the need to model large scale temporal data streams.

Early machine learning literature on the concept drift problem involved mostly computational learning theory tools (Helmbold & Long, 1994; Kuh et al., 1990). Hulten et al. (2001) studied the problem in the context of datamining large scale streams whose distribution change in time. More recently, Forman (2006) studied the concept drift phenomenon in the context of information retrieval in large textual databases. Sharan and Neville (2007) consider the modeling of temporal changes in relational databases and its application to text classification.

Overall, the prevailing techniques have been to train standard methods on examples obtained by filtering the data through a sliding window. Tibshirani and Hastie (1987) developed the local likelihood idea in the statistics community within the context of non-parametric smoothing and regression. More details on local likelihood can be found in (Loader, 1999).

3. The Concept Drift Phenomenon and its Estimation

Formally, the concept drift phenomenon may be thought of as a smooth flow or transition of the joint distribution of a random vector. We will focus on the case of a joint distribution of a random vector X and a random variable Y representing predictor and response variables. We will also restrict our attention to temporal or one dimensional drifts.

Definition 1. *Let X and Y be two discrete random vectors taking values in $\mathcal{X}$ and $\mathcal{Y}$. A smooth temporal drift of X, Y is a smooth mapping from $I \subset \mathbb{R}$ to a family of joint distributions*

$$t \mapsto p_t(x, y) \stackrel{\text{def}}{=} p_t(X = x, Y = y).$$

By restricting ourselves to discrete random variables we can obtain a simple geometrical interpretation of concept drift. Denoting the simplex of all distributions over the set S by

$$\mathbb{P}_S \stackrel{\text{def}}{=} \left\{ r \in \mathbb{R}^{|S|} \;:\; \forall i \, r_i \geq 0, \; \sum_{i=1}^{|S|} r_i = 1 \right\} \quad (1)$$

we have that Definition 1 is equivalent to a smooth parameterized curve in the simplex $\mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$.

The drift in the joint distribution can be decomposed in several ways. The first decomposition $p_t(x, y) = p_t(x|y)p_t(y)$ is useful for generative modeling and the second decomposition $p_t(x, y) = p_t(y|x)p_t(x)$ is useful for conditional modeling. In the generative case we will focus on modeling $p_t(x|y)$ since modeling $p_t(y)$ is typically an easier problem due to its lower dimensionality (in most cases involving text documents $|\mathcal{Y}| \ll |\mathcal{X}|$). In the case of conditional modeling, we focus on modeling $p_t(y|x)$ and we ignore the drift in the marginal $p_t(x)$ since it is irrelevant for discriminative tasks.

In both cases we assume that our data is a set of time-stamped labeled documents sampled from $p_t(x, y)$ where the time points t are sampled from a distribution $g(t)$. If g is a continuous density, the number of samples at time t, denoted by N_t, is no greater than 1 with probability 1. In practice, however, we allow N_t to be larger than 1 in order to account for the discretization of time. We thus have the data

$$D = \{(x_{tj}, y_{tj}) : t \in I \subset \mathbb{R}, j = 1, \ldots, N_t\} \quad (2)$$

where the time points are sampled from $g(t)$ and $(x_{tj}, y_{tj}) \sim p_t(x, y)$.

To illustrate these concepts in the context of the RCV1 dataset, we display in Figure 2 the total number of words per day (left) and the total number of documents per day (right) corresponding to the most popular category in RCV1. As is evident from the right panel, $g(t)$ is a highly non-uniform density corresponding to varying amount of news content in different dates.

It is easy to come up with two simple solutions to the problem of concept drift modeling. The first solution, called the extreme global model, is to simply ignore the temporal drift and use all of the samples in D regardless of their time stamp. This approach results in a single global model $\hat{p}$ which serves as an estimate for the entire flow $\{p_t, t \in I\}$ effectively modeling the concept drift as a degenerate curve equivalent to a stationary point in the simplex. The second simple alternative, called the extreme local model, is to model p_t using only data sampled from time t i.e. $\{(x_{tj}, y_{tj}) : j = 1, \ldots, N_t\}$. This alternative decomposes the concept drift estimation into a sequence of disconnected estimation problems.

The extreme local model has the benefit that if the individual estimation problems are unbiased, the estimation of the concept drift is unbiased as well. The main drawback of this method is the high estimation variance resulting from the relatively small number of daily samples N_t used to estimate the individual models. Furthermore, assuming D is finite we can only estimate the drift in the finite number of time points appearing in the dataset D (since we have no training data for the remaining time points). On the other hand, the extreme global model enjoys low variance since it uses all data points to estimate p_t. Its main drawback is that it is almost always heavily biased due to the fact that samples from one distribution p_{t_1} are used to estimate a different distribution p_{t_2}.

It is a well known fact that the optimal solution in terms of minimizing the mean squared estimation error usually lies between the extreme local and extreme global models. An intermediate solution can trade-off increased bias for reduced variance and can significantly improve the estimation accuracy. Motivated by this principle, we employ local smoothing in forming a local version of the maximum likelihood principle which includes as special cases the two extreme models mentioned above. The intuition behind local smoothing in the present context is that due to the similarity between p_t and $p_{t+\epsilon}$, it makes sense to estimate p_t using samples from neighboring time points $t + \epsilon$. However, in contrast to the global model the contribution of points sampled from $p_{t+\epsilon}$ towards estimating p_t should decrease as ϵ increases.

4. Local Likelihood and Concept Drift

The local likelihood principle extends the ideas of non-parametric regression smoothing and density estimation to likelihood-based inference. We concentrate on using the local likelihood principle for estimating $p_t(x|y)$ and $p_t(y|x)$ which are described next.

4.1. Local Likelihood for n-Gram Estimation

We apply local likelihood to the problem of estimating $p_t(x|y)$ by assuming the naive Bayes assumption i.e. that $x|y$ is generated by a multinomial distribution or its n-gram extensions. Assuming documents contain words belonging to a finite dictionary of size V, the naive Bayes assumption may be stated as

$$p_t(x|y) \propto \prod_{w \in V} \theta_w^{c(w,x)}, \qquad \theta \in \mathbb{P}_V \quad (3)$$

where $c(w, x)$ represents the number of times word w appears in document x. Similarly, the n-gram model extends naive Bayes (3) by considering n-order Markov dependency. The naive Bayes and n-gram are a mainstay of statistical text processing (Manning & Schutze, 1999) and usually lead to accurate language modeling, especially when appropriate smoothing is used (Chen & Goodman, 1998). For notational simplicity we consider the problem of estimating $p_t(x)$ rather than the equivalent $p_t(x|y)$ and we concentrate on naive Bayes i.e. 1-gram. Extending the discussion to n-grams with $n > 1$ is relatively straightforward and is omitted due to lack of space.

Applied to the concept drift problem, the local log-likelihood at time t is a smoothed or weighted version of the loglikelihood of the data D in (2) with the amount of smoothing determined by a non-negative smoothing kernel $K_h : \mathbb{R} \to \mathbb{R}$

$$\ell_t(\eta|D) \stackrel{\text{def}}{=} \sum_{\tau \in I'} K_h(t - \tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}\,; \eta). \quad (4)$$

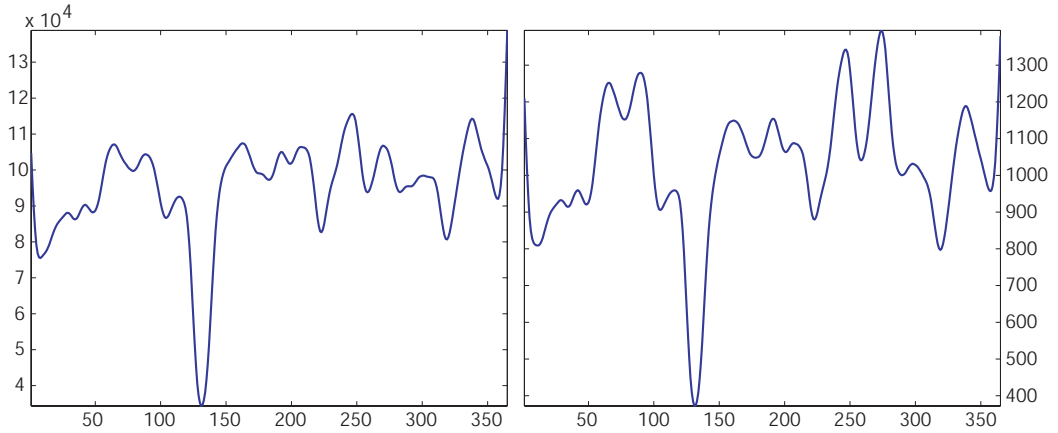

Figure 2. Total number of words per day (left) and documents per day (right) for the most popular category in RCV1. The displayed curves were smoothed to remove sampling noise.

We assume that the kernel function is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ reflecting its spread and satisfying the relation $K_h(r) = h^{-1}K(r/h)$ for some $K : \mathbb{R} \to \mathbb{R}$ referred to as the base kernel form. We further assume that K has bounded support and $\int u^r K(u)\, du < \infty$ for $r \leq 2$. Wand and Jones (1995) provide more details on the formal requirements of a smoothing kernel.

Three popular kernel choices are the tricube, triangular and uniform kernels, defined as $K_h(r) = h^{-1}K(r/h)$ where the $K(\cdot)$ functions are respectively

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \quad (5)$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \quad (6)$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}}. \quad (7)$$

The uniform kernel is the simplest choice and leads to a local likelihood (4) equivalent to filtering the data by a sliding window i.e. $\hat{\theta}_t$ is computed based on data from adjacent time points with uniform weights. Unfortunately, it can be shown that the uniform kernel is suboptimal in terms of its statistical efficiency or rate of convergence to the underlying distribution (Wand & Jones, 1995). Surprisingly, the triangular kernel has a higher statistical efficiency than the Gaussian kernel and is the focus of our experiments in this subsection. We use the tricube kernel in the next subsection.

The scale parameter h is central to the bias-variance tradeoff. Large h represents more uniform kernels achieving higher bias and lower variance. Small h represents a higher degree of locality or lower bias but higher variance. Since $\lim_{h\to 0} K_h$ approaches Dirac's delta function and $\lim_{h\to\infty} K_h$ approaches a constant function the local log-likelihood (4) interpolates between the loglikelihoods of the extreme local model and the extreme global model mentioned in Section 3 as h ranges from 0 to $+\infty$.

Solving the maximum local likelihood problem for each t provides an estimation of the entire drift $\{\hat{\theta}_t : t \in \mathbb{R}\}$ with $\hat{\theta}_t = \arg\max_{\eta\in\Theta} \ell_t(\eta|D)$. In the case of the naive Bayes or n-gram model we obtain a closed form expression for the local likelihood maximizer $\hat{\theta}_t$ as well as convenient expressions for its bias and variance. In general, however, there is no closed form maximizer and iterative optimization algorithms are needed in order to obtain $\hat{\theta}_t = \arg\max_{\eta\in\Theta} \ell_t(\eta|D)$ for all t.

We denote the length of a document in (2) by $|x_{tj}| \stackrel{\text{def}}{=} \sum_{v\in V} c(x_{tj}, v)$ and the total number of words in day t in (2) by $|x_t| \stackrel{\text{def}}{=} \sum_{j=1}^{N_t} |x_{tj}| = \sum_{v\in V}\sum_{j=1}^{N_t} c(v, x_{tj})$. We assume that the length of documents x_{tj} is independent of t and is drawn from a distribution with expectation λ.

Under the above assumptions, the local likelihood (4) of the naive Bayes model becomes

$$\ell_t(\eta|D) = \sum_{\tau\in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \sum_{w\in V} c(w, x_{\tau j}) \log \eta_w$$

where $\eta \in \mathbb{P}_V$. The local likelihood has a single global maximum whose closed form is obtained by setting to 0 the gradient of the Lagrangian

$$0 = \frac{1}{[\hat{\theta}_t]_w} \sum_{\tau\in I} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j}) + \lambda_w$$

to obtain

$$[\hat{\theta}_t]_w = \frac{\sum_{\tau \in I} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j})}{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|}. \quad (8)$$

The estimator $\hat{\theta}_t$ is a normalized linear combination of word counts where the combination coefficients are determined by the kernel function and normalized by the number of words in different days. We note that $\hat{\theta}_t$ in (8) is different from a weighted averaging of the relative frequencies $c(w, x_{\tau j}) / \sum_{w'} c(w', x_{\tau j})$.

We distinguish between two fundamental scenarios for predicting the drift θ_t.

Offline scenario: The goal is to estimate the drift $\{\theta_t : t \in \mathbb{R}\}$ given the entire dataset D. In this case we will consider symmetric kernels $K(r) = K(-r)$ which will achieve an increased convergence rate of $\hat{\theta}_t \to \theta_t$ as indicated by Proposition 2.

Online scenario: The goal is estimate a model for the present distribution θ_t using training data from the past i.e. a dataset whose time stamps are strictly smaller than t. This corresponds to situations where the data arrives sequentially as a temporal stream and at each time point a model for the present is estimated using the available stream at that time. We realize this restriction by constraining K to satisfy $K(r) = 0, r \leq 0$.

As with other statistical estimators, the accuracy of $\hat{\theta}_t$ may be measured in terms of its mean squared error $\mathsf{E}\,(\hat{\theta}_t - \theta_t)^2$ which decomposes as the sum of the squared bias and variance of $\hat{\theta}_t$. Examining these quantities allow us to study the convergence rate of $\hat{\theta}_t \to \theta$ and its leading coefficient .

Proposition 1. *The bias vector* $\mathsf{bias}(\hat{\theta}_t) \stackrel{\text{def}}{=} \mathsf{E}\hat{\theta}_t - \theta_t$ *and variance matrix of* $\hat{\theta}_t$ *in* (8) *are*

$$\mathsf{bias}(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|\,(\theta_\tau - \theta_t)}{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|} \quad (9)$$

$$\mathsf{Var}(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h^2(t-\tau)|x_\tau|\,(diag(\theta_\tau) - \theta_\tau \theta_\tau^\top)}{\left(\sum_{\tau \in I} K_h(t-\tau)|x_\tau|\right)^2} \quad (10)$$

where $diag(z)$ *is the diagonal matrix* $[diag(z)]_{ij} = \delta_{ij} z_i$.

Proof. The random variable (RV) $c(w, x_{\tau j})$ is distributed as a sum of multivariate Bernoulli RVs, or single draws from multinomial distribution. The expectation and variance of the estimator are that of a linear combination of iid multinomial RVs. To conclude the proof we note that for $Y \sim \mathrm{Mult}(1, \theta)$, $\mathsf{E}\,Y = \theta$, $\mathsf{Var}(\theta) = \mathrm{diag}(\theta) - \theta\theta^\top$. □

Examining Equations (9)-(10) reveals the expected dependency of the bias on h and θ_t. The contribution to the bias of the terms $(\theta_\tau - \theta_t)$, for large $|\tau - t|$, will decrease as h decreases since the kernel becomes more localized and will reduce to 0 as $h \to 0$. Similarly, for slower drifts, $\|\theta_\tau - \theta_t\|, t \approx \tau$ will decrease and reduce the bias.

Despite the relative simplicity of Equations (9)-(10), it is difficult to quantitatively capture the relationship between the bias and variance, the sample size, h, λ, and the smoothness of θ_t, g. Towards this goal we derive the following asymptotic expansions.

Proposition 2. *Assuming (i)* θ, g *are smooth in* t, *(ii)* $h \to 0, hn \to \infty$, *(iii)* $g > 0$ *in a neighborhood of* t, *and (iv) document lengths do not depend on* t *and have expectation* λ, *we have in the offline case*

$$\mathsf{bias}(\hat{\theta}_t|I) = h^2 \mu_{21}(K) \left(\dot{\theta}_t \frac{g'(t)}{g(t)} + \frac{1}{2}\ddot{\theta}_t \right) + o_P(h^2) \quad (11)$$

$$\mathsf{Var}(\hat{\theta}_t|I) = \frac{\mu_{02}(K)}{(nh)g(t)\lambda}(diag(\theta_t) - \theta_t\theta_t^\top) + o_P((nh)^{-1})$$

and in the online case

$$\mathsf{bias}(\hat{\theta}_t|I) = h\mu_{11}(K)\dot{\theta}_t + o_P(h) \quad (12)$$

$$\mathsf{Var}(\hat{\theta}_t|I) = \left(\frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda} \right)(diag(\theta_t) - \theta_t\theta_t^\top) + \frac{\mu_{12}(K)}{n\lambda g(t)}(diag(\dot{\theta}_t) - \dot{\theta}_t\theta_t^\top - \theta_t\dot{\theta}^\top) + o_P((nh)^{-1})$$

where $\mu_{kl}(K) \stackrel{\text{def}}{=} \int u^k K^l(u)\,du$ *is assumed to be finite and* $\dot{\theta}_t$ *is the vector* $[\dot{\theta}_t]_i = \frac{d}{dt}[\theta_t]_i$.

The proof is somewhat similar to the derivation of the asymptotic bias and variance of the Nadaraya-Watson local regression (Wand & Jones, 1995) and is omitted due to space limitations. The notation $g_n \xrightarrow{p} f$ represents convergence in probability of g_n to f i.e. $\forall \epsilon > 0, P(|g_n - f| > \epsilon) \to 0$, and $g_n = o_P(f_n)$ represents $g_n / f_n \xrightarrow{p} 0$.

Corollary 1. *Under the assumptions in Proposition 2, and in particular* $h \to 0, nh \to \infty$, *the estimator* $\hat{\theta}_t$ *is consistent i.e.* $\hat{\theta}_t \xrightarrow{p} \theta_t$ *in both the offline and online settings.*

Proposition 2 specifies the conditions for consistency as well as the rate of convergence. In particular, the bias of online kernels converges at a linear rather than quadratic rate. In either cases, the estimator is biased and inconsistent unless $h \to 0, n \to \infty$ and $nh^{-1} \to$

∞. Expressions (11)-(12) reveal the performance gain associated with a slower drift and sampling density g indicated by $\dot{\theta}_t$ and $g'(t)$ and with more (represented by n) and longer (represented by λ) documents.

Figure 3 displays the RCV1 per-word test set loglikelihood for the online and offline scenarios as a function of the (triangular) kernel's bandwidth. As expected, offline kernels performs better than online kernels with both achieving the best performance for a bandwidth approximately 25 which corresponds to a support of 25 days in the online scenario and 50 days in the offline scenario. Note that in addition to obtaining higher accuracy than the global model corresponding to $h \to \infty$, the local model enjoys computational efficiency as it ignores a large portion of the training data.

A central issue in local likelihood modeling is selecting the appropriate bandwidth h. A practical solution is to use cross validation or some other automatic bandwidth selection mechanism. On RCV1 data, the performance of such cross validation schemes is very good and the estimated bandwidth possesses test set loglikelihood that is almost identical to the optimal bandwidth (see Figure 4, left).

Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates (see Figure 4, right). A time-dependent cross validation procedure may be used to approximate the time-dependent optimal bandwidth which performs slightly better than the fixed-date cross validation estimator. Note that the accuracy with which the cross validation estimator approximates the optimal bandwidth is lower in the time-dependent or varying bandwidth situation due the fact that much less data is available in each of the daily cross validation problems.

From a theoretical perspective, the asymptotic bias and variance can be used to characterize the optimal bandwidth and study its properties. Minimizing the (offline) leading term of sum of component-wise MSE with respect to h we obtain the bandwidth estimator

$$\hat{h}_t^5 = \frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j\,/2\right)^2}. \quad (13)$$

As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}\|$ increases. Intuitively this makes sense since in these cases the variance decreases and bias either increases or stays constant. In practice, $\dot{\theta}_t, \ddot{\theta}_t$ may vary significantly with time which leads to the conclusion that a single bandwidth selection for all t may not perform adequately. These changes are illustrated in Figure 5 (left) which demonstrates the temporal change in the gradient norm.

Perhaps more interesting than the dependency of the optimal bandwidth on $n, \lambda, \dot{\theta}_t, \ddot{\theta}_t$ is its dependency on the time sampling distribution $g(t)$. Equation (13) reveals an un-expected non-monotonic dependency of the optimal bandwidth in $g(t)$. The dependency, expressed by $\hat{h}_t \propto (\sum_{j=1}^V (c_{1j}/\sqrt{g(t)} + c_{2j}\sqrt{g(t)})^2)^{-1/5}$ is illustrated in Figure 6 (left) where we assume for simplicity that c_{1j}, c_{2j} do not change with j resulting in $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$. The key to understanding this relationship is the increased asymptotic bias due to the presence of the term $g'(t)/g(t)$ in Equation (11). Intuitively, the variations in $g(t)$ expressed by $g'(t)$ introduce a bias component which alters the otherwise monotonic role of the optimal bandwidth and bias-variance tradeoff. Since $g(t)$ is highly non-uniform (as illustrated in Figure 2), this dependency of $\hat{h}_t$ on $g(t)$ is likely to play a significant role.

We finally point out that different words w have different parameters $[\theta_t]_w$ and parameter derivatives $[\dot{\theta}_t]_w$ which indicates that it is unlikely that a single bandwidth will work best for all words. Frequent words are likely to benefit more from narrow kernel smoothing than rare words which almost never appear. As a result, a lower bandwidth should be used for frequent words while a high bandwidth should be used for rare words. A systematic investigation of these topics is beyond the scope of this paper.

4.2. Local Likelihood for Logistic Regression

Often, the primary goal behind modeling the drift is conditional modeling i.e. predicting the value of y given x. In this case, drift modeling should focus on estimating the conditional $p_t(y|x)$ since modeling the marginal $p_t(x)$ becomes irrelevant. In contrast to the modeling of the conditional by Bayes rule $p_t(y|x) \propto p_t(x|y)p_t(y)$ described in the previous section, we explore here direct modeling of $\{p_t(y|x) : t \in I\}$ using local likelihood for logistic regression.

By direct analogy to Equation (4) the conditional local likelihood estimator $p_t(y|x)$ is the maximizer of the locally weighted conditional loglikelihood

$$\ell_t(\eta|D) = \sum_{\tau\in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(y_{\tau j}|x_{\tau j};\eta) \quad \eta \in \Theta.$$

As in the generative case, the kernel parameter h balances the degree of the kernel's locality and controls the bias-variance tradeoff.

Denoting by $f(x)$ the vector of relative frequencies in the document x, the logistic regression model

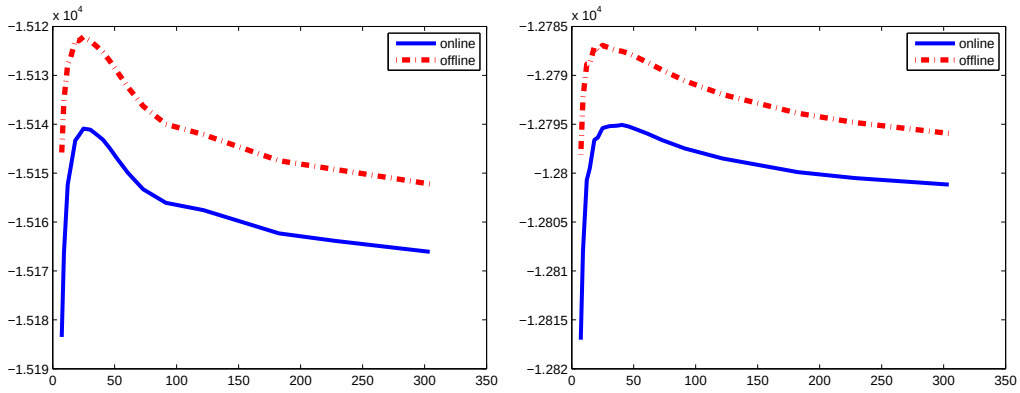

Figure 3. Per-word log-likelihood of held out test set as a function of the triangular kernel's bandwidth for the two largest RCV1 categories (CCAT (left) and GCAT (right)). In all four cases, the optimal bandwidth seems to be approximately 25 which indicates a support of 25 days for the online kernels and 50 days for the offline kernels.

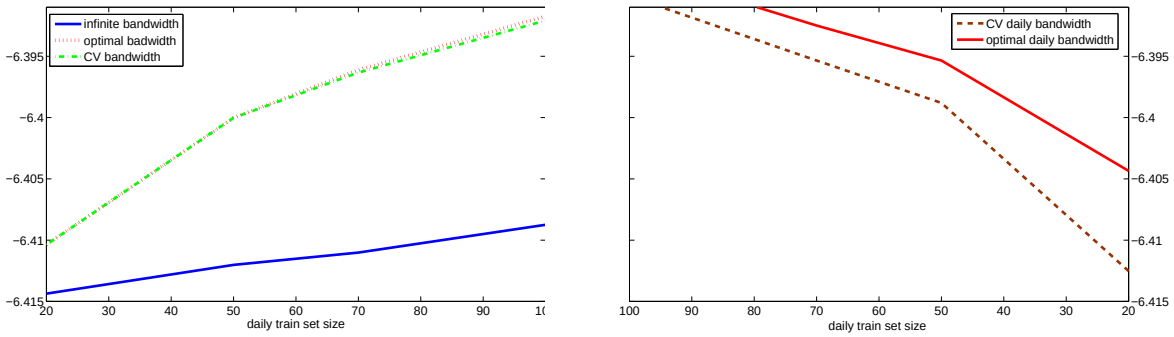

Figure 4. Per-word log-likelihood over held-out test set for various bandwidths as a function of the daily training set size. Left: The extreme global model corresponding to $h \to \infty$ performs worst. Selecting the bandwidth by cross validation results in an accurate estimate and test-set loglikelihood almost identical to that of the optimal slope. Right: Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates.

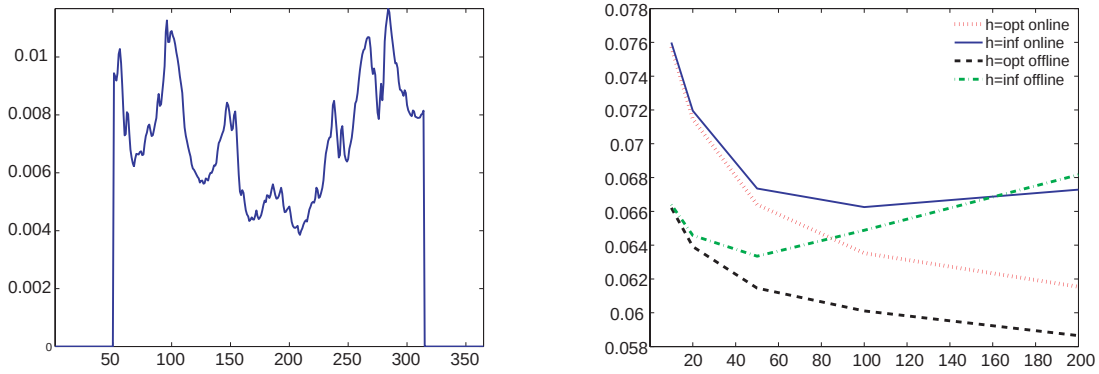

Figure 5. Left: Estimated gradient norm for the most popular category in RCV1 as a function of t. The derivatives were estimated using local smoothing. To avoid running into boundary effects we ignore the first and last 50 days. Right: Classification error rate over a held-out test set for the local logistic regression model as a function of the train set size.

$\log \frac{p(1|x;\theta_t)}{1-p(1|x;\theta_t)} = \langle \theta_t, f(x) \rangle$, $\theta \in \mathbb{R}^V$ leads to the following local conditional likelihood

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + e^{-y_{\tau j}\langle x_{\tau j}, \eta \rangle}\right).$$

In contrast to the naive Bayes model in the previous section, the local likelihood does not have a close form maximizer. However, it can be shown that under mild conditions it is a concave problem exhibiting a single global maximum (for each t) (Loader, 1999). Most of the standard iterative algorithms for training logistic regression can be modified to account for the local weighting introduced by the smoothing kernel. Moreover, recently popularized regularization techniques such as the penalty $c\|\eta\|^q, q = 1, 2$ may be added to the local likelihood to obtain a local regularized version equivalent to maximum posterior estimation.

Figure 5 (right) displays classification error rate over a held-out test set for local logistic regression as a function of the train set size. The classification task was predicting the most popular class vs the second most popular class in RCV1. The plots in the figure contrast the performance of the online and offline tricube kernels with optimal and infinite bandwidths, using L_2 regularization. The local model achieved a relative reduction of error rate over the global model by about 8%. As expected, the online kernel generally achieve worse error rates than the offline kernels. In all the experiments mentioned above we averaged over multiple random samplings of the training set to remove sampling noise.

5. Discussion

A large number of textual datasets such as emails, webpages, news stories, etc. contain time stamped documents. For such datasets, considering a drifting rather than a stationary distribution is often appropriate. The local likelihood framework provides a natural extension for many standard likelihood models to the concept drift scenario. As the drift becomes more noticeable and the data size increases the potential benefits of local likelihood methods over their extreme global or local counterparts increase.

In this paper we illustrate the drift phenomenon and examine the properties of the local likelihood estimator including the asymptotic bias and variance tradeoff and optimal bandwidth. Experiments conducted on the RCV1 dataset demonstrate the validity of the local likelihood estimators in practice and contrast them with more standard non-local alternatives.

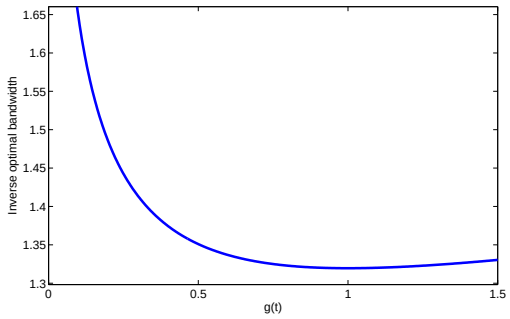

Figure 6. Inverse of the optimal bandwidth derived from Equation (13) as a function of $g(t)$: $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$ (we take $c_1 = c_2 = 1$). The graph show the non-monotonic dependency between $\hat{h}^{\text{opt}}$ and $g(t)$.

References

Chen, S., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modelling* (Technical Report). Harvard university.

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. *Proc. of the ACM SIGIR Conference.*

Helmbold, D. P., & Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning, 14.*

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *Proc. of the ACM SIGKDD Conference.*

Kuh, A., Petsche, T., & Rivest, R. L. (1990). Learning time-varying concepts. *Advances in Neural Information Processing Systems, 3.*

Lewis, D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research, 5.*

Loader, C. (1999). *Local regression and likelihood.* Springer.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The federalist.* Addison Wesley.

Mulligan, G. (1999). *Removing the spam: Email processing and filtering.* Addison Wesley.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales. *Proc. of the ACL conference.*

Sharan, U., & Neville, J. (2007). Exploiting time-varying relationships in statistical relational models. *Proc. of the joint WebKDD and SNA-KDD workshop.*

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *J. of the American Statistical Association, 82.*

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing.* Chapman and Hall/CRC.