Generalized Expectation Criteria

Andrew McCallum, Gideon Mann, Gregory Druck
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{mccallum,gmann,gdruck}@cs.umass.edu

DRAFT

Technical Report UM-CS-2007-60

August, 2007

Abstract

This note describes *generalized expectation* (GE) criteria, a framework for incorporating preferences about model expectations into parameter estimation objective functions. We discuss relations to other methods, various learning paradigms it supports, and applications that can leverage its flexibility.

1 Introduction

The parameters of probabilistic models are often set by maximum (aposteriori) likelihood estimation, moment matching, or the maximum entropy principle. In many common cases, these three parameter estimation methods are actually equivalent. However, each provides its own perspective on the parameter estimation problem; each provides different types of flexibility; and each lends itself different classes of variations outside the equivalence class.

This note describes *generalized expectation* (GE) criteria, which in some cases also falls into the same equivalence class, and, similarly, provides yet another different perspective, a different flexibility, and useful variations outside the common equivalence class. Below we describe GE; we illustrate how it can be used as an augmentation to, or replacement for, traditional parameter estimation methods such as maximum likelihood; and we outline GE's many use-cases in supervised learning, semi-supervised learning, semi-supervised clustering, transfer learning and more.

A generalized expectation criterion is a term in a parameter estimation objective function that expresses some preference about the model's expectations of variable values. That is, a generalized expectation criterion is a function, G, that maps to a scalar the model's predicted distribution over the values of some other function, f, of the model variables; and this scalar is added to the objective function used for parameter estimation.

For example, the function G may express a preference that variable expectations (with f being the identity function) have a certain value by returning a distance between the model's expected value and some target value. The target value would typically be obtained from some external knowledge source, such as training data or prior knowledge. But there are many alternative formulations. For instance, G may be not based on distance to a single target value, but on a smooth hinge loss function.

Many traditional parameter estimation methods can be described as preferences about model expectations. (Some of these are related below.)

The non-tradtional "generalized" nature of our proposal that we explore here is that (1) in factor graphs, we may express preferences about expectations on variable subsets that are not in one-to-one correspondence with the variable subsets participating in the parameterized factors of the model; for example, we can express preferences on a subset of model factors, or on marginal distributions larger than model factors; (2) we may calculate and combine expectations conditioned on different circumstances, such as different training data with different properties; (3) the supervised "training signal," whether that be a target expectation, or, more generally, the shape of the score function, G, may come not just from labeled training data, but from any source, including other tasks or human prior knowledge.

In fact, in most of our experiments, the GE terms do not provide enough contraints themselves to be a consistent estimator. In other words, GE is under-specifying the parameters—there may be many different parameter settings that are equally preferred by the GE criteria. Here we rely on combining GE with other objectives. Alternatively, note that GE criteria could also *over-specify* parameter preferences, by, for example, in a sequence model, expressing expecation targets on higher-order Markov dependencies than are captured by the parameters. In this case, the expectations will probably not be matched exactly, but we may be able to leverage conditioning on the entire input sequence to good effect.

One of the key benefits of GE is that it provides a way for humans to directly express preferences to the parameter estimator naturally and easily using the language of *expectations*, rather than the often complex and counter-intuitive language of the *parameters*. Rather than being forced to speak the language of the parameters—which often have complex relations to the model's behavior, and which often interact with each other in subtle ways—people can instead speak the language of the data, the language of desired model outputs, unconstrained by model structure and unconcerned about the complexities of the model parameterization..

Most of our experiments thus far have been in the application of GE to semi-supervised learning. Here unlabeled data is combined with limited supervision, provided by the human trainer in the form of expected prior label distributions (Mann & McCallum, 2007), or class associations for *features* (rather than instances) (Druck et al., 2007b). We find that GE performs better than several alternative semi-supervised methods, and provides better accuracy given the same amount of labeling effort. In some cases, GE with labeled features matches the accuracy of instance labeling with less than one tenth the wall-clock labeling time.

In the next section we define generalized expectation criteria more formally. In the later sections we give examples of GE's use-cases and capabilities; for many of these we have preliminary experimental results published elsewhere.

2 Definition

Let X be some set of variables, with assignments denoted $\mathbf{x} \in \mathcal{X}$. Let θ be the parameters of some model that defines a probability distribution over X, $p_\theta(X)$.

The expectation of some function $f(X)$ according to the model is

$$E_\theta[f(X)] = \sum_{\mathbf{x} \in \mathcal{X}} p_\theta(\mathbf{x}) f(\mathbf{x}),$$

where $f(\mathbf{x})$ is an arbitrary function of the variables $\mathbf{x}$ producing some scalar or vector value. This function of course may depend only on a subset of the variables in $\mathbf{x}$.

Naturally expectations may also be conditioned on certain variable value assignments, for example, when performing "conditional probability training" of some model. In this case the variables are partioned into "input" variables X and "output" variables Y. A set of assignments to input variables (training data instances)

$\tilde{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, ...\}$ may be provided, and the conditional expectation is then

$$E_\theta[f(X,Y)|\tilde{\mathcal{X}}] = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\mathbf{x}\in\tilde{\mathcal{X}}} \sum_{\mathbf{y}\in\mathcal{Y}} p_\theta(\mathbf{y}|\mathbf{x}) f(\mathbf{x},\mathbf{y}).$$

For simplicity, however, in most of the remainder of the paper we include notation for the non-conditional case—the addition of the conditional being straightforward.

Definition: A *generalized expectation* (GE) criteria is a function, G, that takes as an argument the model's expectation of $f(X)$, and returns a scalar, which is added as a term in the parameter estimation objective function:

$$G(E_\theta[f(X)]) \rightarrow \mathbb{R}.$$

In some cases G might be defined based on a distance to some "target value" for $E_\theta[f(X)]$. Let $\tilde{f}$ be the target value, and let $\Delta(\cdot,\cdot)$ be some distance function. In this case, G might be defined:

$$G_{\tilde{f}}(E_\theta[f(X)]) = -\Delta(E_\theta[f(X)], \tilde{f}).$$

As described thus far, GE is quite generic, and encompasses several other traditional parameter estimation methods as special cases. The three main degrees of flexibility which we leverage in a non-traditional way are:

1. A GE criterion is specified independently of the parameterization. In traditional parameter estimation methods for factor graphs there is a one-to-one correspondence between the subsets of variables employed in each parameterized factor of the model and the subsets of variables on which expectations are calculated for the objective function. In GE, each of these subsets may be selected independently. The use of this flexibility is discussed further in section 4.2.

2. Different conditional GE criteria need not all condition on the same circumstances—they can condition on different data sets or different combinations of data sets. The use of this flexibility is discussed further in section 4.1.

3. The supervised "training signal," whether that be a target expectation, or, more generally, the shape of the score function, G, may come not just from labeled training data, but from any source, including other tasks or human prior knowledge.

Thus a GE criterion may be specified independently of the parameterization, and independently of choices of any conditioning data sets. Note also that a GE criterion may operate on some arbitrary subset of the variables in $\mathbf{x}$. Again, the functions f may be defined such that the expection yields moments of the distribution $p_\theta(X)$, or any other arbitrary expectation. The scoring function G and the distance function Δ may be based on information theory, or be arbitrary functions.

GE terms may be used as the sole components of the parameter estimation objective function, or they may be used in conjunction with other terms. Examples of such combinations appear in below.

Naturally, GE may be applied to many different learning paradigms in which objective functions are used, including joint/generative learning, unsupervised learning, conditional/discriminative learning, supervised learning, learning with hidden variables, structure learning, and others. Some examples are discussed below.

3 Some Relations to Other Methods

GE is closely related to several established methods of parameter estimation.

GE is most similar to the **method of moments**, but, again, with non-traditional aspects. Moments in probability spaces are expected values of variables raised to powers (such as mean, $E[x^1]$, and variance, $E[x^2]$). Moment matching estimates parameters by solving equations that equate moments of the model with moments of a training data set. The traditional method of moments could be seen as a special case of GE in which the functions f yield the moments of the distribution on X, and G is based on a distance function that has its miminum when its two arguments are equal, and furthermore it is possible to solve the equivalence.

In GE we may take expectations not of variables raised to powers, but of arbitrary functions. But more significantly, GE expresses arbitrary scalar preferences, not moment equality equations to be solved. (For example, these scalar preferences may be calculated as distance from the model's expectation to a target expectation, or as an arbitrarily-shaped score function of the model's expectation, including hinge-loss or even multi-modal functions.) These scalar preferences allow GE to be combined with other parameter estimation methods in a multi-criteria objective—be that traditional maximum likelihood, or simply a prior on parameters.

Maximum likelihood is a special case of GE in which G is the negative cross entropy between the empirical distribution on X and the model's distribution on

X. In other words, the function f is the "vector indicator" function,[1] $\tilde{f}(X)$ is the empirical distribution of the vector indicator function f applied to X in the training data $\tilde{\mathcal{X}}$ (the elements of the vector $\tilde{f}(X)$ sum to 1), and $G_{\tilde{f}}$ is the negative cross entropy between the elements of $E_\theta[f(X)]$ and $\tilde{f}(X)$. Thus, $E_\theta[f(X)] = \sum_{\mathbf{x}\in\tilde{\mathcal{X}}} p_\theta(\mathbf{x})f(\mathbf{x})$ and $\tilde{f}(X) = (1/|\mathcal{X}|)\sum_{\mathbf{x}\in\tilde{\mathcal{X}}} f(\mathbf{x})$ and we have

$$G_{\tilde{f}} = -\sum_{i=1...|\mathcal{X}|} \tilde{f}(\mathbf{x})_i \log(E_\theta[f(X)]_i),$$

where subscripts i index into the dimensions of the indicator vector.

Furthermore, in *structured* probabilistic models, represented as **factor graphs**, GE's correspondence to maximum likelihood can also be expressed in terms of the factor graph structure. Rather than calculating the *joint* expectations over all model variables at once, we can rather calculate expectations over the subsets of variables participating in each parameterized factor. That is, if we minimize the cross-entropies between (a) the marginal distribution of the variables in each parameterized factor of the model, and (b) the empirical distribution of the same variables, this also yields the maximum likelihood parameters, by the Hammersley-Clifford-Besag theorem. (Of course, a significant additional flexibility that GE provides is the ability to express various preferences for expectation values on variable subsets that *do not* correspond to any factor in the model parameterization.)

When the parameter estimation objective function consists of both some GE terms and model's log-likelihood of some training data, the GE terms may be thought of as a **regularizer or type of prior**.[2] Here the objection function, $\mathcal{O}$, is

$$\mathcal{O} = \sum_{\mathbf{x}} \tilde{p}(x)\log(p_\theta(x)) + G(E_\theta[f(X)]).$$

For example, when $f(X)$ is $\log p(x)$, then $E_\theta[f(X)]$ is the negative entropy of the model's distribution over X, and the GE term may then express a prior preference for $p_\theta(X)$ distributions that are close to uniform, as would a zero-parametered Dirichlet prior for multinomial θ. Further discussion of GE and its relation to priors appears in section 5.8.

Maximum entropy is a special case in which the parameter estimation objective function consists of both the entropy of $p_\theta(X)$, as well as GE terms $G_{\tilde{f}}$, where

[1] A vector indicator function is a function that takes a value $\mathbf{x} \in \mathcal{X}$ and returns a vector of length $|\mathcal{X}|$ containing zeros everywhere except at a position in the vector uniquely associated with the value $\mathbf{x}$, where it contains a 1.

[2] Although it is strange to call it a prior, since it is calculated in terms of expectations on model variables, not directly on parameter values; furthermore, in the conditional training case, the GE criterion model expectation would depend on the input data $\tilde{\mathcal{X}}$.

the $\tilde{f}$s are the constraints and the $f(X)$s are the functions necessary to yield the corresponding expectations from the model, and $G_{\tilde{f}}$ insists on exactly matching expectations and constraints, perhaps modulo the typical L2 or L1 prior. Unlike traditional maximum entropy, however, GE can naturally express constraints that have no corresponding parameter, and could apply to models not in the exponential family.

Note that in the maximum entropy framework, we begin with some statements about desired expectations ("constraints"), then through the application of the maximum entropy principle and Lagrange multipliers we arrive inextricably at a certain resulting model structure and parameterization. All constraints have parameters, and all parameters have constraints—there is a one-to-one correspondence. By contrast, in GE, we start with some arbitrary model, and then estimate its parameters through preferences about some of its expectations—which may not be in one-to-one correspondence with parameters, for example, just a subset of the parameters. We make extensive use of this flexibility in several use-cases below.

4 Discussion of Degrees Flexibility

Rather than seeing traditional methods as special cases of GE, it is more interesting to see what non-traditional approaches come to mind when exploring the flexibility of GE.

4.1 Expectations conditioned on different data sets

When maximizing conditional likelihood, the distribtion over the conditioned variable Y is taken as the empirical distribution over some available data sets $\tilde{\mathcal{X}}$. These data sets may be used in the same way when calculating expectation $E_\theta[f(Y|\tilde{\mathcal{X}})]$ in GE.

However, in GE is the clear that we could use different such input datasets in different GE criterion terms. This flexibility is useful when we have available multiple different datasets with different properties, such as different missing data, or different source distributions. Several possible applications, including semi-supervised learning and transfer learning, are briefly described below.

4.2 Different coverage of parametric factors and constraints-expectations

Traditionally in factor graphs or in maximum entropy models there is a one-to-one correspondence between the subsets of variables appearing the parameterized factors of the model and the subsets of variables appearing in the constraints used for training. In GE it is clear that we have the flexibility to break this correspondence because GE terms are defined separately from the model.

GE training expectations could be expressed in finer grainularity than the parameters. That is, $f(X)$ could employ a subset of variables larger than any subset on which a parameterized factor is defined. In this case the model would not have individual parameters capable of tuning these expectations, but it could use the interactions with neighboring factors to try to minimize these GE terms.

GE training expectations could also be expressed in coarser grainularity than the parameters. That is, $f(X)$ could employ a subset of variables smaller than any subset on which a parameterized factor is defined. In this case, the model has a higher degree of expressivity than the constraints require. Other GE constraints or other non-GE terms in the objective function may express preferences for the full degree of expressivity; however, the coarser-grained GE constraints may be used to leverage a different set of trainining data in which the full granularity is not available, (*e.g.* a case of transfer learning). Furthermore, coarse-grained GE terms may simply express preferences for certain marginal distributions not explicitly enforced previously.

Of course, GE training expectations can also be expressed on just a fraction of all the factors (variable subsets) in the model parameterization. In some cases, there may be no term in the objective function (or no term, beside a parameter prior) that expresses a preference about certain factors. In this case, the parameters in those factors will be set in whatever way most helps satisfy the GE terms that are present. This case is related to hidden variables—variables not present in the training data—but the GE view on this phenomena is flexible in a different way because it is expressed in terms of *factors* missing from the training criteria, not variables.

GE's overall flexibility in coverage of parametric factors and constraints-expectations suggests discussion about "two factor graphs": one factor graph describing the variable subsets employed in the parameterization of the model, and another factor graph describing the variable subsets employed in the constraints and expectations used for training. For training such a model, we would need inference in the "cross product" of the two factor graphs.

4.3 Flexible Supervision

In maximum likelihood parameter estimation, the parameters are set to give highest probability to the model generating some observed set of data. In maximum apriori estimation, the parameters are set to those that give high value to the product of this data likelihood and a prior distribution on parameters. Usually this prior is selected to be "non-informative."[3]

It is our belief that too often machine learning is performed *tabula rasa*. The model parameters are estimated from labeled training, but without the benefit of domain knowledge from a human expert. Human experts have helpful knowlege about what good solutions look like, but there have not been intuitive, convenient ways to express this knowledge in the parameter estimation objective function.

GE, however, makes it easy for a human expert to directly express highly specific supervision signals—input to training not based on labeled training data. A human expert can make natural statements such as "I would expect that a professor's home page would have more hyperlinks to students' pages than would another student's," or "I would expect that 70% or more of the documents containing the word ice would be about ICEHOCKEY instead of BASEBALL," and these can be directly translated into GE terms that guide parameter estimation.

Without GE, the primary way that domain experts inject their knowledge is through the selection of model structure and feature selection. But this still leaves a significant parameter estimation problem. Furthermore, although crucial, model structure selection and feature selection may be a difficult medium of communication between a domain expert and a machine learning expert because they are often quite technical considerations that domain experts may have difficulty understanding, *e.g.* ("Should we use a Gaussian or a log-normal distribution? I don't know; what is a log-normal distribution again?").

Supervision concerning GE expectation terms can come from a variety of sources. As described above, humans can express these directly. But also they could come from other related tasks. Recently we have been performing transfer learning experiments in which GE-style supervision from one task helps estimate the parameters for another task. This is described further below.

[3] Although it may also be designed to be "informative." However, this is relatively rare, perhaps because, as described above, priors are specified "in the language of parameter space," and it is difficult for humans to speak this language given the complexities of the model and the interactions among its parts.

5 Use Cases

5.1 Application to semi-supervised learning

In semi-supervised learning we perform learning for a supervised task in which only a small portion of training data is labeled, but we have available a large data set with labels missing. GE expectations may be calculated separately using these different data sets. That is, we may condition on the unlabeled data $\tilde{\mathcal{X}}$ when calculating model expectations, $E_\theta[f(x, y)]$ for some of the GE terms.

Note in particular that GE terms may express constraints on only a fraction of the factors in the model, (for example, only on the features appearing in the labeled data, or only on features over which prior knowledge is expressed), however, other factors (features appearing only in the unlabeled data) will have their parameters usefully set, as described in section 4.2. Based on co-occurrences, these parameters on unlabeled-only features will be set help to satisfy the GE terms.

Mann and McCallum (2007) describe an application of GE to semi-supervised learning in which the GE terms express preferences about marginal class distributions. These marginal class distributions can often be provided by human domain experts, or may be robustly estimated from labeled data, even when labeled data is far too sparse to estimate the many parameters on input-feature/label affinities.

Let $\tilde{\mathbf{f}} = \tilde{p}(Y)$ be some target distribution over class labels, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_i^n \vec{I}(y_i)$ where $\vec{I}$ denotes the vector indicator function on labels $y \in \mathcal{Y}$ and n is the number of output variables[4]. The expectation of $f(\mathbf{x}, \mathbf{y})$ is then the model predicted distribution over labels. A GE term might simply be defined as the negative[5] KL-divergence between these distributions

$$-D\left(\tilde{\mathbf{f}}, \frac{1}{|\tilde{\mathcal{X}}|}\sum_{\mathbf{x}\in\tilde{\mathcal{X}}}\sum_{\mathbf{y}\in\mathcal{Y}} p_\theta(\mathbf{y}|\mathbf{x}) f(\mathbf{x}, \mathbf{y})\right). \tag{1}$$

5.2 Application to semi-supervised clustering

Closely related to semi-supervised learning is semi-supervised clustering, in which unsupervised clustering is guided by limited human input.

[4] For example, for a classification task $n = 1$, while for a sequence labeling task n would be the length of the sequence.

[5] Throughout this section we assume we aim to maximize the objective function.

GE terms can straightforwardly be added to objective functions for unsupervised learning with latent variables. These GE terms may express various human preferences about certain expectations, including guiding some latent variables to certain positions, guiding more or fewer latent variables to cover certain regions of space, encouraging sparseness properties through GE terms that score second moments, or enforcing certain constraints on model predictions (as in the applications described in Sections 5.1 and 5.3).

As one possibility, suppose we have a "prototype" instance $\mathbf{x}'$ for cluster y. We can encourage the model for cluster y to give high probability to similar instances by including a GE term

$$\sum_{\mathbf{x} \in \tilde{\mathcal{X}}} p_\theta(\mathbf{x}|y) \mathrm{sim}(\mathbf{x}, \mathbf{x}'), \tag{2}$$

where sim is some similarity function, for example cosine similarity.

5.3 Application to semi-supervised learning with feature labeling

GE also makes very natural an under-explored paradigm for semi-supervised learning—one in which the limited supervision is in the form of "feature labeling" (or more generically "expectation labeling") rather than "instance labeling." In text classification, for example, rather than examining and labeling documents, human labelers instead simply indicate a relatively small set of words that are positively correlated with each class. This human input can be readily translated into constraints on model expectations for certain feature-label combinations. Druck et al. (2007b) explore this scenario, showing that learning with labeled features gives much better accuracy given the same amount of labeling effort.

Specifically, suppose we have prior knowledge about some feature of an input variable $f_i(x)$. For example $f_i(x)$ may indicate that x is a specific word, or that it matches some regular expression. Let $\tilde{\mathbf{f}} = \tilde{p}(Y|f_i(x)=1)$ be the target distribution over labels conditioned on feature $f_i(x)$ being present. For simplicity, here we consider problems with a single output variable y. Let $f(\mathbf{x}, y) = \frac{1}{C_i} \vec{I}(y) \sum_j^n f_i(x_j)$, where $C_i = \sum_j^n f_i(x_j)$, so that the expectation is the model predicted class distribution when $f_i(x)$ is present. We encourage these distributions to agree using negative KL-divergence

$$-D\left(\tilde{\mathbf{f}}, \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\mathbf{x} \in \tilde{\mathcal{X}}} \sum_{y \in \mathcal{Y}} p_\theta(y|\mathbf{x}) f(\mathbf{x}, y)\right). \tag{3}$$

5.4 Application to constraints beyond the model factors

Since GE terms are specified indepedently of the model or its factorization, it is natural to consider terms expressing preferences that are more specific than the model itself can represent.

For example, the model may be a linear-chain CRFs with first-order Markov dependencies, but a GE objective function may express preferences about second-order Markov statistics. When using generative training (for example, an HMM), this may not have much effect, but in a conditionally-trained model (a CRF), there is freedom to include features of the input from an arbitrarily-sized window—one of higher Markov over than the factors on the output—and the parameters can use the long-range view of the input to try to satisfy these preferences.

Specifically, consider a linear chain CRF with parameterized factors of the form $\Psi_t(\mathbf{x}, \mathbf{y}) = \exp(\sum_i \theta_i f_i(x_t, y_t, y_{t+1}))$. We could specify a GE term that scores the model expectation of a function that looks at an additional input and output variable, such as $f(x_t, x_{t+1}, y_t, y_{t+1}, y_{t+2})$, without including a corresponding parameter θ in the model.

5.5 Application to transfer learning and distantly labeled data

Transfer learning applies when we have labeled data for a task that is related but not identical to the target task. In addition we may have some limited labeled data for the target task. GE applies very well to this scenario because it can help us manage two key difficulties in transfer learning.

First, when we can identify which features from the related task apply robustly also to the target task, we can emphasize those in the GE terms. If we learn mappings from "related features" to "target features" we can create GE terms that leverage the mapping.

For example, let $f_i(x)$ be some feature of an input variable x that is relevant to both tasks. Let $\tilde{\mathcal{S}}$ be labeled source domain data and $\tilde{\mathcal{T}}$ be unlabeled target domain data. We define the reference expectation $\tilde{\mathbf{f}} = \tilde{p}_{\tilde{\mathcal{S}}}(Y|f_i(x) = 1)$, which is the observed class distribution when feature $f_i(x)$ is present in the labeled source data $\tilde{\mathcal{S}}$. We let $f(\mathbf{x}, y) = \frac{1}{C_i}\vec{I}(y)\sum_j^n f_i(x_j)$ (as in Section 5.3). The GE term encourages agreement between the source and target expectations

$$-D\left(\tilde{\mathbf{f}}, \frac{1}{|\tilde{\mathcal{T}}|}\sum_{\mathbf{x}\in\tilde{\mathcal{T}}}\sum_{y\in\mathcal{Y}} p_\theta(y|\mathbf{x}) f(\mathbf{x}, y)\right). \tag{4}$$

Second, sometimes the data for the related task may be missing so much context, that the model does not directly apply to it. In this case, this "distantly labeled data" can be external knowledge source used to robustly estimate the constraints $\tilde{f}(X)$. For example, a lexicon of city names is lacking the context of surrounding natural language so crucial to the information extraction model, but the lexicon (in conjunction with other lexicons and background knoweldge) can be used to set GE constraints $\tilde{f}(X)$ on the affinity between certain words and the LOCATION label. Druck et al. (2007a) describe experiments with this approach.

5.6 Application to active learning

Active learning occurs when not all training supervision is available at the beginning of training; rather, it is solicited by the machine to the human during the learning process—hopefully in such a way as to minimize the human effort required to reach a certain level of accuracy.

Typically active learning has consisted of the machine solicting labels for instances it carefully chooses. GE makes it natural to solicit additional types of feedback, such as "labeled features" as described in section 5.3 above.

As described in the next subsection, GE also opens up several avenues for communication between the human and the model which may be useful for active learning.

5.7 GE as a language for safe communication with the model

Although there has been much research in active learning, it is rarely deployed in practice. The more common human-machine interaction cycle when using machine learning to build a new system is (1) label data, (2) train and test an initial model, (3) perform error analysis, (4) make adjustments to the model, adding new features, adding new parameters, (5) go back to step 2, repeating as necessary.

In our experience in steps (3) and (4) there are numerous times when we might see some repeated egregious error, wonder how the model could make such a mistake, and wish we could "reach into the model parameterization" and directly jack up or down the value of some parameter.

But doing this could be very dangerous. Parameter values in factor graphs are delicately balanced against each other. Manually tuning some parameter after training could have unintended consequences. Humans can rarely speak safely in the langage of parameter values.

GE, however, allows the human to interact or "speak to the model" in terms of *expectations* rather than *parameter values*. Expectations are expressed in terms of the data, about which the human will likely have better intuitions than the effects of the model parameters.

If during error analysis we see some repeated egregious error, we may add some new GE constraint (which is more than just adding a feature—it also adds preferences for certain values), and thus nudge the model in the intended direction, knowing that further training will safely preserve the delicate balance among parameters.

5.8 Expressive language for "priors"

In traditional maximum aposteriori parameter estimation desiterata for the likelihood term typically come from training data and are expressed in terms of feature functions, while desiterata for the parameter prior term typically are expressed in terms of parameter values, *e.g.* preferring parameter values close to zero.

GE-style "priors" are expressed in terms of model expectations. This provides a more natural method for incorporating domain knowledge into the objective function than *informative priors* or *subjective likelihood* [M. Lavine 2007], because domain knowledge is often naturally expressed in terms of expectations; human-performed attempts to translate such domain knowledge directly into individual parameter value preferences is fraught with difficulty, since parameters and their effects typically interact with each other in complex, subtle ways. However, the mathematics of GE would perform this translation automatically.

Charles Sutton has pointed out that G could represent a *probability distribution* over expected values of $f(X)$. Distance functions to a target can also represent varying degrees of preference for different expected values, but human experts may find it easier to express preferences in probability density functions than in distances.

This point is also related to empirical Bayes. David Blei points to work in empirical Bayes by Brad Efron in which expectations on the data are used to determine certain priors.

6 Conclusions

Generalized expectation (GE) criteria provide desirable flexibility that will be useful in several types of learning tasks.

Our claim is *not* that GE is a fundamentally new statistical estimator. We are merely advocating for GE as a unexplored perspective that naturally suggests creative solutions to some important problems.

We expect that there is a significant amount of related work. We will update this note with further related work in the future.

We are excited about the future possibilities of GE, and have already begun experimentation with some of the approaches described above. In addition to work in semi-supervised learning with label regularization (Mann & McCallum, 2007), and semi-supervised learning with feature labeling (Druck et al., 2007b), recent preliminary experiments in domain adaptation and active learning are yielding positive results.

References

Druck, G., Mann, G., & McCallum, A. (2007a). Leveraging existing resources using generalized expectation criteria. *NIPS 2007 Workshop on Learning Problem Design*.

Druck, G., Mann, G., & McCallum, A. (2007b). *Reducing annotation effort using generalized expectation criteria* (Technical Report 2007-62). University of Massachusetts, Amherst.

Mann, G., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. *ICML*.