Optimization Algorithms for Data Analysis

Stephen Wright

University of Wisconsin-Madison

Banff, March 2011

Outline

Review some optimization tools of possible relevance in cancer treatment and data analysis.

- Learning from Data: SVM classification, regularized logistic regression
- Sparse optimization (with group sparsity)
- Nonlinear optimization for biological objectives

1. Learning from Data

Learn how to make inferences from data.

Related Terms: Data Mining, Machine Learning, Support Vector Machines, Classification, Regression, Kernel Machines.

Given numerous examples ("training data") along with the correct inferences for each example, seek rules that can be used to make inferences about *future* examples.

Among many possible rules that explain the examples, seek simple ones.

- Expose the most important features of the data.
- Easy and inexpensive to apply to future instances.
- More generalizable to the underlying problem - don't over-fit to the particular set of examples used.

Rich source of sparse and regularized optimization problems.

Binary Labels

Have feature vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^m$ (real vectors) and binary labels $y_1, y_2, \ldots, y_n = \pm 1$.

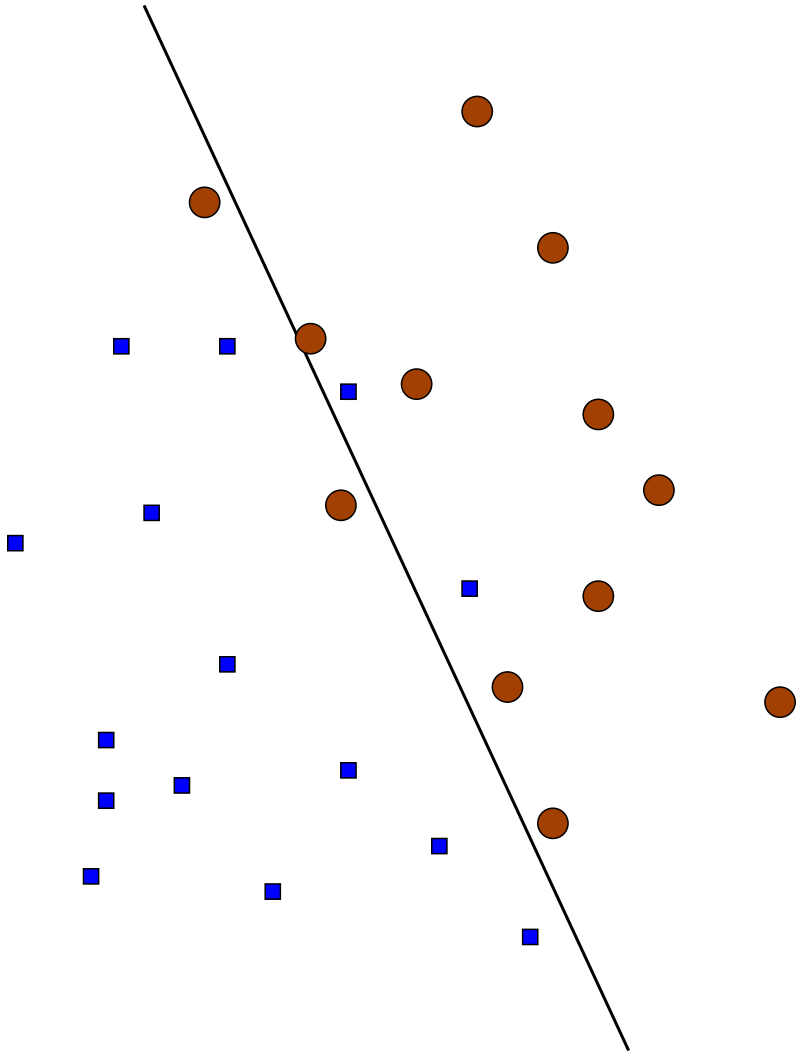
Seek rules that predict the label on future examples $x \in \mathbb{R}^m$.

Classification: Learn a function $f : \mathbb{R}^m \to \{-1, +1\}$ such that the predicted label is $f(x)$.

Odds: Learn functions $p_+ : \mathbb{R}^m \to [0, 1]$ such that $p_+(x)$ is the chance of x having label $+1$, and $p_-(x) := 1 - p_+(x)$ the chance of label -1.

Many variants, e.g. multiple classes (> 2), some or all examples unlabelled.

Linear Support Vector Machines (SVM) Classification



- Seek a hyperplane $w^T x + b$ defined by coefficients (w, b) that separates the points according to their classification:

$$w^T x_i + b \geq 1 \Rightarrow y_i = 1, \qquad w^T x_i + b \leq -1 \Rightarrow y_i = -1$$

(for most training examples $i = 1, 2, \ldots, n$).

- Penalized formulation: for some $\lambda > 0$, solve

$$\min_{(w,b)} \frac{\lambda}{2} w^T w + \frac{1}{m} \sum_{i=1}^{m} \max\left(1 - y_i[w^T x_i + b], 0\right).$$

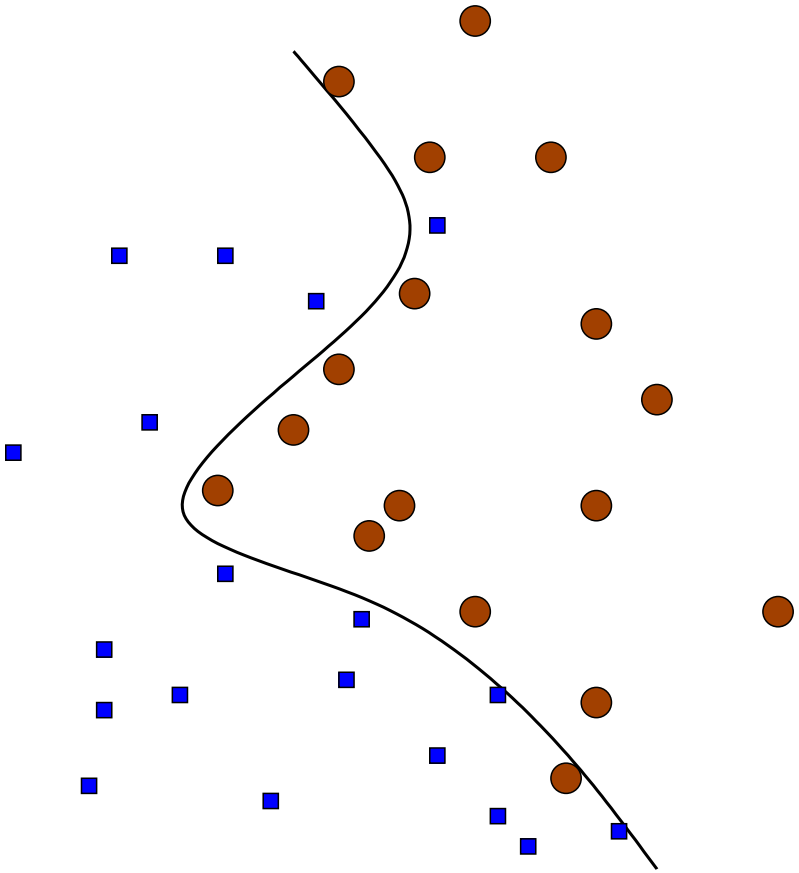
Term i in summation is 0 if point i is correctly classified, positive otherwise.

- Dual:

$$\max_{\alpha} e^T \alpha - \frac{1}{2} \alpha^T \tilde{K} \alpha \;\text{ s.t. }\; \alpha^T y = 0, \;\; 0 \leq \alpha \leq \frac{1}{\lambda m} \mathbf{1},$$

where $y = (y_1, y_2, \ldots, y_m)^T$, $\tilde{K}_{ij} = y_i y_j x_i^T x_j$.

Nonlinear Support Vector Machines



Nonlinear SVM

To get a *nonlinear* classifier, map x into a higher-dimensional space $\phi : \mathbb{R}^n \to \mathcal{H}$, and do linear classification in $\mathcal{H}$ to find $w \in \mathcal{H}$, $b \in \mathbb{R}$.

When the hyperplane is projected back into $\mathbb{R}^n$, gives a nonlinear surface (often not contiguous).

In "lifted" space, primal problem is

$$\min_{(w,b)} \frac{\lambda}{2} w^T w + \sum_{i=1}^{m} \max\left(1 - y_i[w^T \phi(x_i) + b], 0\right).$$

By optimality conditions (and a representation theorem), optimal w has the form

$$w = \sum_{i=1}^{m} \alpha_i y_i \phi(x_i).$$

Kernel

By substitution, obtain a finite-dimensional problem in $(\alpha, b) \in \mathbb{R}^{m+1}$:

$$\min_{\alpha,b} \frac{\lambda}{2}\alpha^T \Psi \alpha + \frac{1}{m}\sum_{i=1}^{m} \max\left(1 - \Psi_{i\cdot}\alpha - y_i b, 0\right),$$

where $\Psi_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$. WLOG can impose bounds $\alpha_i \in [0, 1/(\lambda m)]$.

Don't need to define ϕ explicitly! Instead define the kernel function $k(s,t)$ to indicate distance between s and t in $\mathcal{H}$.

Implicitly, $k(s,t) = \langle \phi(s), \phi(t) \rangle$.

The Gaussian kernel $k^G(s,t) := \exp(-\|s-t\|_2^2/(2\sigma^2))$ is popular.

Thus define $\Psi_{ij} = y_i y_j k(x_i, x_j)$ in the problem above.

Logistic Regression

Parametrize p_+ by a vector z: Seek to learn a weight vector $z \in \mathbb{R}^m$ such that the following functions give the odds of a new feature vector x belonging to class $+1$ and -1, resp.:

$$p_+(x;z) = \frac{1}{1+e^{z^T x}}, \qquad p_-(x;z) = \frac{1}{1+e^{-z^T x}}.$$

(Note that $p_+ + p_- \equiv 1$.) Denote $L_+ := \{i \,|\, y_i = +1\}$, $L_- := \{i \,|\, y_i = -1\}$.

- For $x_i \in L_+$, want $z^T x_i \ll 0$, so that $p_+(x_i;z) \approx 1$.
- For $x_i \in L_-$, want $z^T x_i \gg 0$, so that $p_-(x_i;z) \approx 1$.

ℓ_1 regularization

Negative, scaled a posteriori log likelihood function is

$$\mathcal{L}(z) = -\frac{1}{n}\left[\sum_{i\in L_-} \log p_-(x_i; z) + \sum_{i\in L_+} \log p_+(x_i; z)\right]$$
$$= -\frac{1}{n}\left[\sum_{i\in L_-} z^T x_i - \sum_{i=1}^{n} \log(1 + e^{z^T x_i})\right].$$

LASSO-Pattersearch: Seek a solution z with few nonzeros, by adding a regularization term $\tau\|z\|_1$:

$$\min_z \; T_\tau(z) := \mathcal{L}(z) + \tau\|z\|_1.$$

Smaller $\tau \Rightarrow$ more nonzeros in solution z.

Application: Eye Study: Interacting Risk Factors

- W. Shi, G. Wahba, S. J. Wright, K. Lee, R. Klein, and B. Klein, "LASSO-Patternsearch algorithm with application to opthalmology data," *Statistics and its Interface* 1 (2008), pp. 137-153. Code: `http://pages.cs.wisc.edu/~swright/LPS/`

Beaver Dam Eye Study. Examined 876 subjects for myopia.

- 7 risk factors identified: gender, income, juvenile myopia, cataract, smoking, aspiring, vitamin supplements.
- Bernoulli model: Chose a cutpoint for each factor, assign 1 for above cutpoint and 0 for below.
- Examine all $2^7 = 128$ interacting factors.

The four most significant factors are:

- cataracts (2.42)
- smoker, don't take vitamins (1.11)
- male, low income, juvenile myopia, no aspirin (1.98)
- male, low income, cataracts, no aspirin (1.15)

plus an intercept of -2.84.

A much larger application about genetic risk factors for rheumatoid arthritis also studied ($> 400,000$ variables).

Extensions

Multiple Outcomes. Extensions to multiclass SVM and logistic regression are known.

Regression rather than Classification. Needed when the outcome is not discrete. Support Vector Regression, variable selection in data fitting could be used.

Using the Results to Drive Optimization. Having identified the most important effects, how do we change the way we formulate the optimization problems assciated with treatment planning?

Uncertainty in Data or Outcomes. Can the conclusions be made robust to uncertainty and errors in data or outcomes?

2. Sparse Optimization

Many applications prefer *structured*, *approximate* solutions of optimization formulations, to *exact* solutions.

- Data inexactness does not warrant exact solution;
- Simple solutions may be easier to actuate and easier to understand;
- Avoid "overfitting" to a particular sample of the full data set;
- Extract just the essence of the knowledge contained in the problem specfication, not the less important effects;
- Too much computation needed for an exact solution.

To achieve the desired structure, can modify the problem formulation and the algorithms used to solve it.

("Sparse" refers to the solution vector, not the problem data.)

Applications

- Compressed sensing
- Image deblurring and denoising
- Matrix completion (e.g. netflix prize)
- Low-rank tensor approximations for multidimensional data
- Machine learning e.g. support vector machines.
- Logistic regression and other variable selection problems in computational statistics

Beam selection problems in treatment planning could potentially be modeled and solved with a "group-sparse" formulation — see below.

Regularized Formulations

Directly imposing the desired structure can lead to a computationally difficult problem.

Example: Seek $x \in \mathbb{R}^n$ that approximately minimizes $f(x)$ and has at most r nonzeros. We can model this using e.g. n binary variables to turn components of x on and off. Or solve with a customized branch-and-bound procedure — but both are expensive!

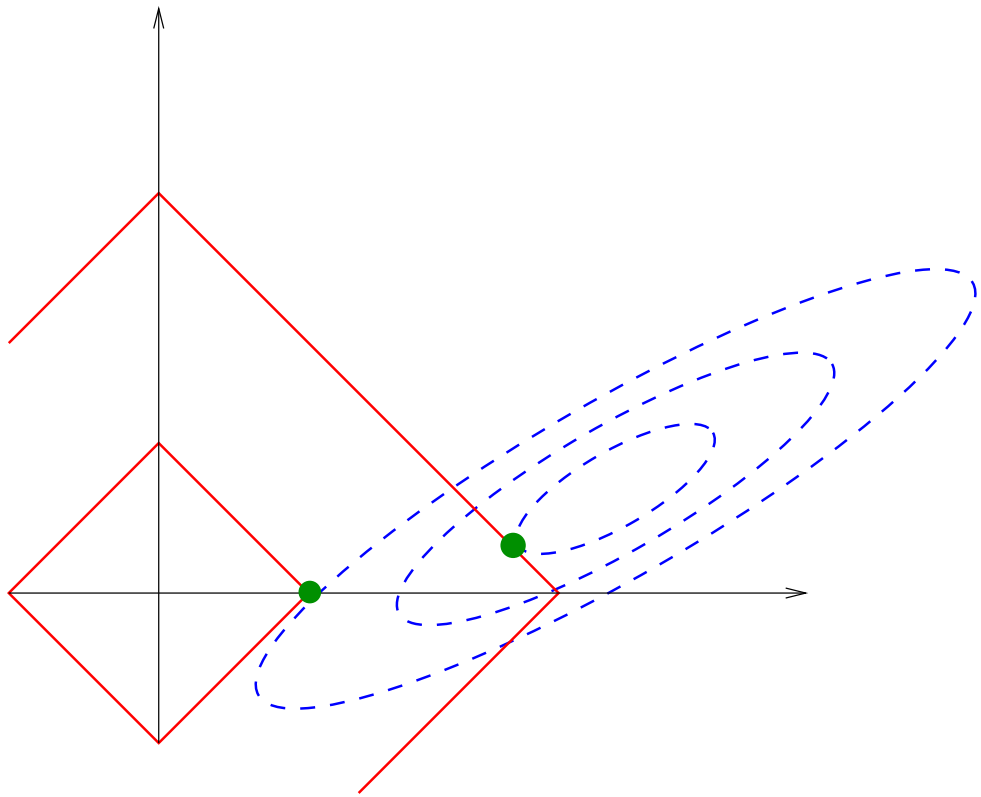
Can instead add a regularization term $P(x)$ to the objective $f(x)$, to promote the kind of structure desired.

P is usually nonsmooth, with "kinks" at values of x with the desired structure. (The kinks "add volume" to the subgradient.)

Example: $P(x) = \|x\|_1$ promotes vectors x with few nonzeros.

$$\min f(x) \text{ s.t. } \|x\|_1 \leq T, \qquad \text{for some } T > 0.$$

Solution has a single nonzero for small T, two for larger T.



An equivalent weighted formulation is

$$\min f(x) + \tau P(x), \qquad \text{for some } \tau > 0.$$

When $P(x) = \|x\|_1$, larger $\tau \Rightarrow$ fewer nonzeros in x.

If we seek approximate minimizers x most of whose components have the form $\pm\sigma$, we use the regularizer $P(x) = \|x\|_\infty$.

Often want solutions for a grid or range of parameters τ, not just one. From this range of values, choose the solution that has the desired sparsity or structure.

Group Regularizers

The regularizer $\|\cdot\|_1$ is ubiquitous and useful — but treats all components of x independently. In some applications, components of x can be arranged naturally into groups.

Denote groups by $[q] \subset \{1, 2, \ldots, n\}$, where $q = 1, 2, \ldots, Q$. Each group is a subvector of x, denoted by $x_{[q]}$, $q = 1, 2, \ldots, Q$. The groups may or may not overlap.

Regularizers that promote group sparsity (turns the $x_{[q]}$ on and off as a group):

$$P(x) = \sum_{q=1}^{Q} \|x_{[q]}\|_2, \qquad P(x) = \sum_{q=1}^{Q} \|x_{[q]}\|_\infty.$$

(Sum-of-ℓ_2 and sum-of-ℓ_∞. Both nonsmooth.)

Beamlet Selection

In treatment planning, x consists of beamlet intensities. Groups may consist of the beamlets from one beam angle. The group regularizers above would thus "select" the appropriate beam angle from among many possibilities, and also assign beamlet weights.

This is an alternative to other beam selection techniques, e.g. column generation, heuristics, binary variables.

For other devices and other treatment planning methodologies, other regularizers may be appropriate.

Solving Regularized Formulations

Many tools and techiques needed:

- Large-scale optimization: gradient projection, optimal first-order, sampled gradient, second-order, continuation, coordinate relaxation, interior-point, ...
- Nonsmooth optimization: cutting planes, subgradient methods, successive approximation, ...
- Duality
- Numerical linear algebra
- Heuristics

Also a LOT of domain-specific knowledge about the problem structure and the type of solution demanded by the application.

Basic Algorithm: Prox-Linear

At iteration k, solve for step d^k:

$$\min_d \nabla f(x^k)^T d + \frac{\alpha_k}{2} d^T d + \tau P(x^k + d),$$

choosing α_k large enough to give descent in the objective $f + \tau P$.

When P is separable (i.e. when groups are disjoint, or when $P(x) = \|x\|_1$), can solve the subproblem cheaply and in closed form.

Enhancements to Prox-Linear

Many enhancements are available in the important special cases $P(x) = \|x\|_1$ and P group-separable.

- Compute step in just a subset $\mathcal{G}_k \subset \{1, 2, \ldots, n\}$ of the components of x. Thus need to evaluate only the $\mathcal{G}_k$ components of ∇f.
- Keep track of "apparently nonzero" component set $\mathcal{A}_k$; periodically take reduced approximate Newton steps on this set. Requires only (approximate) Hessian over the components in $\mathcal{A}_k$.
- Use continuation in τ: Solve first for a large τ (easier problem), then reduce τ and re-solve, using previous solution as a starting point. Repeat as needed.

3. Optimizing Biological Objectives

Biological objectives can have different features from physical, dose-matching objectives.

- Highly nonlinear
- Ill conditioned
- Nonconvex

Sometimes the nonlinearity comes from smoothing of kinks in the objective.

Algorithms need to be able to deal with these features, and also exploit the structure of these objectives.

Experience with EUD

Example: Objectives based on Equivalent Uniform Dose (EUD)[1]

- EUD is a nonlinear, possibly nonconvex function of physical dose distribution on a region. It can capture min, max, or average dose to voxels in the region, depending on parameter settings.
- Used in conjunction with logistic functions to devise penalty functions for "soft" upper or lower bounds to dose in a region.
- Objective combines these penalties. (Unknowns: beamlet weights.)

Solve $\min_{w \geq 0} f(w)$. Use a two-metric projected gradient framework. At each iterate w^k, requires

- calculation of gradient $\nabla f(w^k)$,
- estimation of the active set $\mathcal{A}_k$ of components of w that are "probably nonzero" at the solution;
- estimate of the reduced Hessian $\nabla^2 f(w^k)$ on the components in $\mathcal{A}_k$.

[1] Olafsson, Jeraj, Wright, PMB 2005

Using Reduced Hessian Information

In regularized logistic regression and EUD applications (and many others), use of second-order information on the set of apparently-nonzero components can greatly speed convergence of the method.

- Apparently-nonzero set is often relatively small.
- Reduced Hessian may be much less expensive than full Hessian.
- Hessian is highly structured; may be able to get an approximation cheaply.
- Even when Hessian is ill-conditioned, the Newton equations may be solvable approximately in just a few steps of conjugate gradient (because right-hand side of Newton equations tends to be in the range space of the large-eigenvalue part of the Hessian).

Final Thoughts

Have not covered many other areas of optimization that may be relevant for modern cancer radiotherapy planning.

- Handling uncertainty
- Formulations based on risk functionals
- Complex outcome-based objectives
- Formulations involving DVH constraints
- Assimilating information from multiple scans
- Using feedback to tune doses during the course of treatment.

FIN