CISC 371 Class 36

Summary of Nonlinear Data Analysis

In this course, we have touched on many technical matters related to data analysis. We assumed, as our intellectual foundation, that linear algebra and linear data analysis were familiar to us. We formulated our analysis methods by using an *objective function* that we sought to minimize, recognizing that a local non-global minimizer would most commonly be the result of our computations.

We began by understanding objective functions with a single scalar argument. Although this was a simple model, and relatively uncommon empirically, it helped us to understand key concepts such as:

- Relationships among derivatives
- Separation of search direction and stepsize
- Dynamic stepsize selection, especially by back-tracking
- The utility of linear models in an iterative framework

Before we proceeded to searching in a vector space, we needed some mathematical objects. The derivative of a scalar function with a vector argument was written as a *1-form*; this is a dual of a vector so, if a vector was written as a matrix with one column, then a 1-form was written as a matrix with one row. This notation clarified the use of a gradient and the distinction between a gradient and a vector. We wrote the direction of *steepest descent* for a vector argument as the negated transpose of the gradient.

Next, we studied *unconstrained optimization*. This model used an objective function with a single vector argument. The search process was to iteratively update an estimate of a minimizer argument, using a *search direction* and a stepsize. The basic method was steepest descent; alternative methods included Newton's method and scaling methods, which are descent methods that alter the direction of steepest descent by a transformation that is a symmetric positive semidefinite matrix. Unconstrained optimization was extended to nonlinear least squares, training neural networks, and Tikhonov regularization that altered to objective function with an additional term.

Our final exploration was an introduction to *constrained optimization*. We used Lagrange multipliers to transform a constrained problem into an unconstrained problem that could be solved using descent methods. In some applications, we found that the dual Lagrange formulation was a computationally superior expression of a constrained problem. We explored the support vector machine, or SVM, which incorporated kernel methods and soft margins in the dual formulation.

36.1 Review – Functions With A Scalar Argument

This material includes scalar optimization. The main concepts include:

- Approximation: Taylor series; quadratic fit, with and without derivatives
- Iteration with fixed stepsize; possible difficulties
- Stepsize selection by backtracking; Armijo's method

Each student is expected to be able to apply these concepts to specific problems, such as: finding stationary points; assessing a stationary point for convexity; modest calculations for fixed stepsize and backtracking; explaining likely behavior of a method for a function when you are provided with an initial estimate.

36.2 Review – Functions With A Scalar Argument

This material includes: basic relevant vector calculus; the definition of a gradient as a 1-form; and unconstrained optimization that uses the methods of steepest descent and Newton's method.

The main concepts for basic relevant vector calculus include:

- Definitions of partial derivative $\partial/\partial w_j$, directional derivative $D_{\vec{v}} f(\vec{w}_0)$
- Equation relating directional derivative to partial derivatives
- Definition of gradient 1-form $\underline{\nabla} f$ using partial derivatives
- Jacobian matrix $J_{\vec{f}}(\vec{w}_0)$ of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ at $\vec{w}_0$
- Hessian matrix $\nabla^2 f(\vec{w}^*)$ at stationary point $\vec{w}^*$
- Level curve of a function f at level l, written as $\mathbb{S}_C(f, l)$, is all $\vec{u} \in \mathbb{R}^n$ such that $f(\vec{u}) = l$

The main concepts for optimization using steepest descent include:

- Definition of a descent direction
- Steepest direction is transpose of gradient
- Fixed stepsize and backtracking for steepest descent
- Scaling and Newton's method for iterative solution

Each student is expected to be able to apply these concepts to specific problems, such as: finding stationary points; assessing a Hessian for convexity; modest calculations for fixed stepsize and backtracking; explaining likely behavior of a method for a function and an initial estimate.

36.3 Review – Artificial Neural Networks

The examinable material includes: linear algebra for artificial neural networks; a single artificial neuron; and a neural network that has a single layer of "hidden" artificial neurons.

The main concepts in linear algebra for neural networks include:

- Data vector is transpose of data observation, so $\vec{x} = \underline{x}^T$
- Augmented observation is $[\underline{x}\ 1]$
- Label of data vector $\vec{x}_j$ is $y_j \in \{0\ 1\}$
- Linear response is $u(\vec{w}) = [\underline{x}\ 1]\vec{w}$
- Kronecker product is $A \otimes C$ and vectorization is $\text{vec}(\vec{a})$
- General matrix equation $AWC = M$ can be solved as $[C^T \otimes A]\text{vec}(W) = \text{vec}(M)$

The main concepts in a single artificial neuron include:

- Activation function is $\phi : \mathbb{R} \to \mathbb{R}$, also called the score $z(u) = \phi(u)$
- Sigmoid $\phi_S(u) = 1/(1 + e^{-u})$, ReLU $\phi_R(u) = \max(0, u)$, Heaviside $\phi_H(u) = (u > 0)$
- Derivative of activation is $\psi(u) = \phi'(u)$
- Quantization function $q : \mathbb{R} \to \{0\ \ 1\}$ is usually the Heaviside step function
- Residual error is $r(\vec{w}) = y - \phi(\vec{w})$
- Objective is $f(\vec{w}) = \frac{1}{2}r^2$
- Learning rate η is a fixed stepsize in steepest descent

The main concepts in a layer of "hidden" artificial neurons include:

- Layer of "hidden" neurons; individual weights ${}_1\hat{w}_j$ gathered into the matrix ${}_1W$
- Linear response of the layer is $\underline{u}({}_1W) = \underline{x}\ {}_1W$
- Activation of hidden layer is ${}_1\underline{z}({}_1W) = [\phi({}_1u_1)\ \cdots\ \phi({}_1u_l)]$
- Input to Layer 2 is ${}_2\underline{x} = [{}_1\underline{z}\ \ 1]$; residual error is $r(\vec{w}) = y - {}_2\phi(\vec{w})$
- Gradient of objective function is

$$\begin{aligned}
\underline{\nabla} f_2(\vec{w}) &= [\underline{\nabla} f_2({}_2\vec{w})\ \underline{\nabla} f_2({}_1\vec{w})] \\
&= \left[[{}_2\vec{s}]_2 b\ \ \text{vec}\left({}_1S^T\right)\right] \\
{}_1S &\stackrel{\text{def}}{=} \begin{bmatrix} {}_1\vec{x} \\ 1 \end{bmatrix} \left[{}_1\vec{\psi}({}_1\vec{u}) \odot {}_1\vec{b}\right]^T \\
{}_1\vec{b} &\stackrel{\text{def}}{=} ({}_2b)\,{}_2\vec{w}_{1..n_2}
\end{aligned}$$

36.4 Review – Constrained Optimization

The examinable material includes: Lagrange dual formulation of a quadratic objective with linear equality constraints; the primal formulation of the Lagrange function, and the Lagrange equation, for linear inequality constraints; and the conditions at a KKT point.

The main concepts for the simple dual formulation include:

- Equality constraints included using Lagrange multiplier μ
- Stationarity conditions express $\vec{w}$ in terms of $\vec{\mu}$
- Substitute into primal formulation to derive dual formulation
- Dual formulation is always concave
- For this special case, dual is a simple quadratic equation in $\vec{\mu}$

The main concepts for the primal formulation include:

- Inequality constraints included using Lagrange multiplier λ
- KKT conditions apply for inequality constraints
- Linear inequalities often solved by inspecting KKT conditions
- Simple special cases are in the notes and as exercises in the references

The main concepts for the conditions at a KKT point include:

- Complementary slackness can pick active constraints, which are satisfied at boundary points
- Lagrange multipliers must be non-negative, or $\vec{\lambda} \geq \vec{0}$
- Primal feasibility is satisfying the constraints
- Stationarity is satisfying the Lagrange equation $[\partial \mathcal{L}/\partial \vec{w}]^T = \vec{0}$
- KKT conditions are often inspected to determine a solution to a specific problem

The main concepts for constrained least squares and cross-validation of linear regression include:

- CLS uses hyper-parameter θ as inequality constraint
- Validation trains using all the data; not representative of actual regression results
- K-fold cross-validation randomly divides data into training subsets and testing subsets
- Leave-one-out is a form of k-fold cross-validation
- Cross-validation can be used to select a hyper-parameter value, e.g., θ or λ
- With a hyper-parameter, need three subsets: train, validate, test
- Cross-validation is an empirically useful way of validating regression

36.5 Review – Optimal Linear Separation With The SVM

The examinable material includes: the primal formulation of an SVM classifier with hard margins; the dual formulation of an SVM classifier; soft margins by means of slack variables; using a kernel function to find support vectors and a bias value; and using a kernel function to score a data vector. The main concepts for the SVM include:

- Primal formulation from margin maximization
- Lagrange function from objective and inequality constraints
- Interpreting Lagrange multipliers and/or support vectors
- Dual formulation from the Lagrange equations
- Soft margins as slack variables
- Using a kernel function to find support vectors and a bias value
- Using a kernel function to score a data vector