Convexity, Loss functions and Gradient

Abhishek Kumar

Dept. of Computer Science, University of Maryland, College Park

Oct 4, 2011

- Linear Classifiers: Decision boundaries are linear

$$\hat{y} = \left(\sum_{i=1}^{d} w_i x_i + b\right) = (\mathbf{w} \cdot \mathbf{x} + b), \qquad \text{Prediction: } \operatorname{sign}(\hat{y})$$

$$\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = 0\} \quad \Rightarrow \quad ?$$

$$\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b > 0\} \quad \Rightarrow \quad ?$$

$$\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b < 0\} \quad \Rightarrow \quad ?$$

$$b = 0 \quad \Rightarrow \quad ?$$

Margin: $y\hat{y}$, Prediction Error: if Margin < 0

- Optimization Problem (for learning): (using 0-1 loss)

$$\mathbf{w} = \arg\min_{\mathbf{w}, b} \frac{1}{N} \sum_{n} \mathbb{I}(y_n(\mathbf{w} \cdot \mathbf{x}_n + b) < 0)$$

  - $\mathbb{I}(\cdot)$: 1 if argument is true, else 0
  - NP hard to optimize (exactly, and even approximately)
  - Small change in $\mathbf{x}$ may cause large change in loss

Regularization

- We care about test error (not training error).
- Minimizing training error alone can overfit the training data.
- Regularized Learning (SRM): Tikhonov, Ivanov, Morozov

$$\mathbf{w} = \arg\min_{\mathbf{w},b} \frac{1}{N} \sum_n \mathbb{I}(y_n(\mathbf{w} \cdot \mathbf{x}_n + b) < 0) + \lambda R(\mathbf{w})$$

- Balances empirical loss and *complexity* of the classifier – we prefer *simple*!
- $R(\mathbf{w})$ is a regularizer for linear hyperplanes.
- For computational reasons, we want both loss function and regularizer to be convex.

Convexity

- Convex Sets: A set S is convex if

$$\forall x, y \in S, \quad \alpha x + (1-\alpha)y \in S \qquad (0 \le \alpha \le 1)$$

(line segment joining x and y is contained in S)

$$S = \{x : x^2 > 2\} \qquad \Rightarrow ?$$
$$S = \{\mathbf{x} : \mathbf{x}'\mathbf{x} < 1\} \qquad \Rightarrow ?$$
$$S = \{\mathbf{U} : \mathbf{U}'\mathbf{U} = \mathbf{I}\} \qquad \Rightarrow ?$$

- Operations that preserve convexity of sets:
  - Intersection: S_1, S_2 convex $\Rightarrow S_1 \cap S_2$ convex
  - Affine function $(f(x) = Ax + b)$:

    S convex $\Rightarrow f(S)$ and $f^{-1}(S)$ convex

Convexity

- Convex Function: Function f defined on a convex set is convex if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \qquad (0 < \alpha < 1)$$

Equivalent definition: Function f is convex if its epigraph is a convex set.

$$E = \{(x, \mu) : \mu \in R, f(x) \leq \mu\}$$

(region that lies above the graph of the function f)

- f convex $\Rightarrow -f$ concave

Convexity

How to check convexity?

Domain should be a convex set and one of following three:

- Use definition (chord lies above the function)

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \qquad (0 < \alpha < 1)$$

- For differentiable functions: Function lies above all tangents

$$f(y) \geq f(x) + f'(x)(y-x)$$
$$(f(y) \geq f(x) + \nabla f(x) \cdot (y-x))$$

- For twice differentiable functions: Second derivative is non-negative

$$f''(x) \geq 0 \qquad (\nabla^2 f(x) \succeq 0)$$

(If the above three are strict inequalities, then f is strictly convex.)

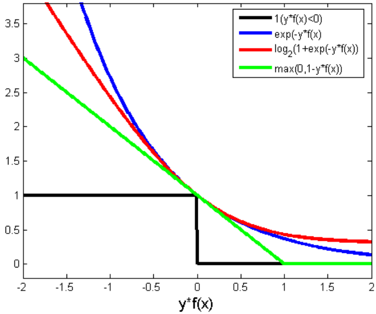
Convexity

- Operations that preserve convexity
  1. Positve scaling and addition: f, g convex$\Rightarrow$ $w_1 f + w_2 g$ convex
  2. Affine function composition: f convex $\Rightarrow$ $f(Ax + b)$ convex
  3. Pointwise maximum: f, g convex$\Rightarrow$ $h(x) = \max(f(x), g(x))$ convex
  4. f convex, g convex non-decreasing $\Rightarrow$ $h(x) = g(f(x))$ convex
  5. f concave, g convex non-increasing $\Rightarrow$ $h(x) = g(f(x))$ convex

- Examples
  - e^x, x, x^2, x^4 etc. $\qquad (x \in R)$
  - Any vector or matrix norm
  - $h(x) = max(|x|, x^2),\ x \in R$ (using property 3 above)
  - $exp(f(x)),\ f$ convex (using property 4 above)
  - $1/log(x),\ x > 1$ (using property 5 above)

Convex Loss Functions

- All of these are convex upper bounds on 0-1 loss.
- Hinge loss: $L(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$
- Exponential loss: $L(y, \hat{y}) = \exp(-y\hat{y})$
- Logistic loss: $L(y, \hat{y}) = \log_2(1 + exp(-y\hat{y}))$



Weight Regularization

Why

- Need weights to be small: ϵ change in input causes $(w \cdot \epsilon)$ change in $\hat{y}$.
- Prediction function should be smooth with respect to input x (should change slowly).
- Regularization is related to generalization:

  Regularized learning is stable (statistically) and stability directly bounds generalization error (or test error).
- We prefer convex regularizers due to computational reasons.

Norm based Regularizers

- ℓ_p Norm: $\quad ||w||_p = \left(\sum_i |w_i|^p\right)^{1/p}, \qquad p \in [1, \infty]$
- $||w||_p$ for $p < 1$ is not a norm. Why?
- Commonly used norm based regularizers in classification: ℓ_1, ℓ_2
- Unit ball in ℓ_p: A set $S = \{w : ||w||_p \leq 1\}$.

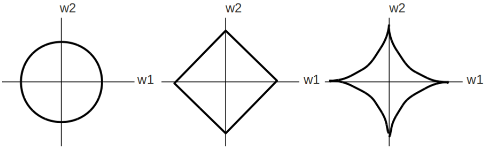


Figure : ℓ_2, ℓ_1, and $\ell_p (p < 1)$ balls in two dimensions

- What will ℓ_∞ ball look like?
- What will ℓ_0 ball look like?

Properties of different regularizers

- Solution often lies on the singularity (corners) of the ball when constrained enough. Why?
- Sparsity inducing norms: obtained for ℓ_p regularizers for $p \leq 1$. Why?
- Rotational Invariance: ℓ_2 norm is invariant to rotations.
- When to use ℓ_1: less number of samples, lot of irrelevant features (feature selection)
- When to use ℓ_2: otherwise. leads to good generalization.

Look Back

- Now we have convex loss functions and convex regularizers. Final objective:

$$\mathcal{L}(\mathbf{w}, b) = \frac{1}{N}\sum_{n} \ell(\hat{y}, y_n) + \frac{\lambda}{2}||w||_2^2$$

- How to minimize and find the solution **w**?

  Gradient Descent.

Gradient

- Gradient of a function f is denoted as ∇f,

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right).$$

- Gradient of a scalar function is a vector with dimension of the domain of function.

- Directional Derivative: Rate of change of function along a direction $\mathbf{v}$.

$$D_v(f) = \nabla f \cdot \mathbf{v}$$

- Function changes (increases) the most along the direction of gradient. Why?

Gradient - geometric interpretation

- Level Sets: For function $f(\mathbf{x})$, the level set containing a point $\mathbf{a}$ is a set of points

$$S = \{\mathbf{x} : f(\mathbf{x}) = f(\mathbf{a})\}$$

  Function admits same value at all points in the level set S.

- Gradient at a point $\mathbf{y}$ is perpendicular to the level set containing $\mathbf{y}$.
- Example: $f(x, y) = \sqrt{x^2 + y^2}$

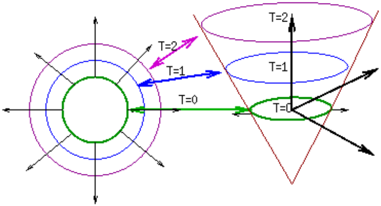


Figure : Level sets of a cone

- How will level sets of function $f(x, y, z) = x^2 + y^2 + z^2$ look like?

Gradient Descent

- Move in the negative direction of gradient to minimize a function.
- Steepest gradient descent: To minimize a function f, start with some initial guess $\mathbf{x}^{(0)}$.

$$\text{Update rule:} \qquad \mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \eta^{(t)} \nabla f(\mathbf{x}^{(t-1)})$$

- Example: Optimizing logistic loss with ℓ_2 regularizer

$$\mathcal{L}(\mathbf{w}, b) = \frac{1}{N} \sum_n \log_2 \left[1 + \exp(-y_n(\mathbf{w} \cdot \mathbf{x}_n + b))\right] + \frac{\lambda}{2} ||w||^2$$