Topics in Machine Learning Midterm 2

2017-04-11

- **Convex optimization**: minimize $f_0(x)$ subject to $f_i(x) \le 0$ and $h_i(x) = 0$
  - f_0 is the **objective function** and *convex*. The problem is *quasiconvex* if f_0 is *quasiconvex*.
  - f_i are **inequality constraint functions** and *convex*.
  - h_i are **equality constraint functions** and *affine*.
- x is a **feasible point** if $x \in D = f_0 \bigcap (\bigcap_i f_i) \bigcap (\bigcap_i h_i)$.
- The optimization problem is said to be **feasible** if there exists at least one feasible point x, and **infeasible** otherwise, i.e. $D = \emptyset$.
- **Optimal value** p^*: $p^* = \infty$ if the problem is *infeasible*. $p^* = -\infty$ if the problem is *unbounded below*.
- x^* is **optimal** if $f_0(x^*) = p^*$.
- x is **locally optimal** if $\exists R > 0$ such that $f_0(z) \ge f_0(x)$ for all feasible z satisfying $\|z - x\|_2 \le R$, i.e. x is optimal within that l2-norm ball of radius R.
- If there exists an optimum x^*, then the optimal value is *attained* or *achieved*, and the problem is solvable. If the **optimal set** (the set of optimum) is empty, then the optimal value is not attained or not achieved.
- Examples of optimal and locally optimal points:
  - $f_0(x) = 1/x$: $p^* = 0$ but not achievable.
  - $f_0(x) = -\log(x)$: $p^* = -\infty$, i.e. unbounded below.
  - $f_0(x) = x \log(x)$: $p^* = -1/e$ is achieved at $x = 1/e$.
  - $f_0(x) = x^3 - 3x$: $p^* = -\infty$, i.e. unbounded below, but a local optimum at $x = 1$.
- **Implicit** v.s. **explicit constraint**: a problem is **unconstrained** if it has *no explicit constraints*.
- **Feasibility problem**:
  - Find a point that satisfies all of the constraints.
  - A special case of convex optimization, where $f_0(x) = 0$.
  - $p^* = 0$ if constraints are *feasible*; any feasible x is optimal.
  - $p^* = \infty$ if constraints are *infeasible*.
- Any locally optimal point of a convex problem is *(globally) optimal*!
  - Proof by contradiction: suppose x is locally optimal, but there exists a feasible y with $f_0(y) < f_0(x)$...

- *Optimality condition* for differentiable f_0: x is optimal iff
  - $\nabla f_0(x)^{\mathrm{T}}(y-x) \geq 0$ for all feasible y
  - Unconstrained problem: $\nabla f_0(x) = 0$
  - Equality constrained problem (subject to $Ax = b$): there exists a ν such that $\nabla f_0(x) + A^{\mathrm{T}}\nu = 0$, i.e. $\nabla f_0(x)$ is the *row space* of A, or the *column space* of A^{T}
  - Minimization over nonnegative orthant, i.e. subject to $x \succeq 0$: $\nabla f_0(x)_i \geq 0$ if $x_i = 0$, or $\nabla f_0(x)_i = 0$ if $x_i > 0$.
- Equivalent convex problems:
  - Eliminating equality constraints
  - Introducing equality constraints
  - Introducing slack variables for linear inequalities
  - Epigraph form
  - Minimizing over some variables

2017-04-18

- **Linear program (LP)**: minimize $c^{\mathrm{T}}x + d$ subject to $Fx \preceq g$ and $Ax = b$
  - *Affine* objective + *Affine* constraints, i.e. feasible set is a polyhedron.
  - Diet problem: minimize $c^{\mathrm{T}}x$ subject to $Ax \succeq b$ and $x \succeq 0$
  - Piecewise-linear minimization: minimize t subject to $a_i^{\mathrm{T}}x + b_i \leq t$, where $i = 1, \ldots, m$
  - Chebyshev center of a polyhedron: maximize r subject to $a_i^{\mathrm{T}}x_c + r\|a_i\|_2 \leq b_i$, where $i = 1, \ldots, m$
- **Quadratic program (QP)**: minimize $\frac{1}{2}x^{\mathrm{T}}Px + q^{\mathrm{T}}x + r$ subject to $Gx \preceq h$ and $Ax = b$
  - *Quadratic* objective + *Affine* constraints, i.e. feasible set is a polyhedron.
  - Least-squares: minimize $\|Ax - b\|_2^2$
  - Linear program with random cost: minimize $\mathbf{E}(c^{\mathrm{T}}x) + \gamma \mathbf{var}(c^{\mathrm{T}}x)$ subject to $Gx \preceq h$ and $Ax = b$
- **Quadratically constrained quadratic program (QCQP)**: minimize $\frac{1}{2}x^{\mathrm{T}}P_0x + q_0^{\mathrm{T}}x + r_0$ subject to $\frac{1}{2}x^{\mathrm{T}}P_ix + q_i^{\mathrm{T}}x + r_i \leq 0$ and $Ax = b$, where $i = 1, \ldots, m$
  - *Quadratic* objective + *Quadratic* inequality constraints + *Affine* equality constraints, i.e. feasible set is intersection of m ellipsoids and an affine set.
- **Second-order cone programming (SOCP)**: minimize $f^{\mathrm{T}}x$ subject to $\|A_ix + b_i\|_2 \leq c_i^{\mathrm{T}}x + d_i$ and $Fx = g$, where $i = 1, \ldots, m$
  - *Affine* objective + *Second-order cone* inequality constraints + *Affine* equality constraints, i.e. feasible set is intersection of m second-order cones and an affine set.

- Convex problem with **generalized inequality constraints**: minimize $f_0(x)$ subject to $f_i(x) \preceq_{K_i} 0$ and $Ax = b$, where $i = 1, \ldots, m$
  - *Convex* objective + *Generalized convex* inequality constraints + *Affine* equality constraints.
- **Conic form problem**: minimize $c^{\mathrm{T}}x + d$ subject to $Fx \preceq_K g$ and $Ax = b$
  - *Affine* objective + *Generalized affine* constraints, i.e. feasible set is a non-polyhedral cone.
- **Semidefinite program (SDP)**: minimize $c^{\mathrm{T}}x + d$ subject to $\sum_{i=1}^{n} F_i x_i + G \preceq_K 0$ and $Ax = b$, where F and G are semidefinite.
  - *Affine* objective + **Linear matrix inequality (LMI)** constraints + *Affine* equality constraints.
  - SDP is more general than LP and SOCP.
  - Eigenvalue minimization: minimize $\lambda_{\max}(A(x))$, where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$
  - Matrix norm minimization: minimize $\|A(x)\|_2 = \lambda_{\max}(A(x)^{\mathrm{T}} A(x))^{1/2}$, where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$
- Vector optimization problem: vector objective $f_0 : \mathbb{R}^n \to \mathbb{R}^q$ minimized w.r.t. proper cone $K \in \mathbb{R}^q$, i.e. f_0 is K-convex.
- **Optimal** and **Pareto optimal** points:
  - Feasible x is *optimal* if $f_0(x)$ is the minimum value.
  - Feasible x is *Pareto optimal* if $f_0(x)$ is a minimal value.
- Multi-objective optimization:
  - If there exists an *optimal* point, the objectives are noncompeting.
  - If there are multiple *Pareto optimal* values, there is a trade-off between the objectives.
  - Regularized least-squares: minimize $(\|Ax - b\|_2^2, \|x\|_2^2)$
- **Lagrangian form** of an optimization problem:
  - $L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$
  - L: $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$
  - λ: Lagrange multiplier associated with $f(x) \leq 0$
  - ν: Lagrange multiplier associated with $h(x) = 0$
- **Lagrange dual function**:
  - $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x))$
  - g: $\mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$
  - g is *concave* no matter the original problem is convex or not.
- **Lower bound property**: if $\lambda \succeq 0$, then $p^* \geq g(\lambda, \nu)$.
- [x] Homework: 4.3, 4.11, 4.23
- Solutions: [Solutions3.pdf]()

2017-05-02

- Lagrange dual and conjugate function: minimize $f_0(x)$ subject to $Ax \preceq b, Cx = d$
  - $g(\lambda, \nu) = -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu$
- **Lagrange dual problem**: maximize $g(\lambda, \nu)$ subject to $\lambda \succeq 0$
  - Finds best lower bound on p^* obtained from Lagrange dual function.
  - *Optimal value* is denoted as d^*
- **Weak duality**: $d^* \leq p^*$
  - Always holds (for convex and nonconvex problems)
- **Strong duality**: $d^* = p^*$
  - Does not hold in general.
  - Conditions that guarantee strong duality in convex problems are called **constraint qualifications**.
- **Slater's condition (constraint qualification)** implies *strong duality* for convex problems:
  - The problem is *strictly feasible*, i.e. x is in the *interior* of D.
  - *Linear inequalities* do not need to hold with strict inequality.
- Least-norm solution of linear equations: minimize $x^T x$ subject to $Ax = b$
  - Maximize $g(\nu) = -\frac{1}{4}\nu^T AA^T\nu - b^T\nu$
- Standard form LP: minimize $c^T x$ subject to $Ax = b, x \succeq 0$
  - $g(\lambda, \nu) = \begin{cases} -b^T\nu \text{ if } A^T\nu - \lambda + c = 0 \\ -\infty \text{ otherwise} \end{cases}$
  - Maximize $-b^T\nu$ subject to $A^T\nu + c \succeq 0$
- Inequality form LP: minimize $c^T x$ subject to $Ax \preceq b$
  - $g(\lambda) = \begin{cases} -b^T\lambda \text{ if } A^T\lambda + c = 0 \\ -\infty \text{ otherwise} \end{cases}$
  - Maximize $-b^T\lambda$ subject to $A^T\lambda + c = 0, \lambda \succeq 0$
- Quadratic program: minimize $x^T Px$ subject to $Ax \preceq b$
  - Maximize $g(\lambda) = -\frac{1}{4}\lambda^T AP^{-1}A^T\lambda - b^T\lambda$ subject to $\lambda \succeq 0$
- Geometric interpretation:
  - *Strong duality* holds if there is a non-vertical supporting hyperplane to A at $(0, p^*)$
  - For convex problem, A is convex, hence has supporting hyperplane at $(0, p^*)$
  - *Slater's condition*: if there exist $(u', t') \in A$ with $u' < 0$, then supporting hyperplanes at $(0, p^*)$ must be non-vertical.
- **Complementary slackness**:
  - Assume *strong duality* holds, x^* is *primal optimal*, (λ^*, ν^*) is *dual optimal*.
  - x^* minimizes $L(x, \lambda^*, \nu^*)$

- $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \ldots, m$, i.e. $\begin{cases} \lambda_i^* > 0 \rightarrow f_i(x^*) = 0 \\ f_i(x^*) < 0 \rightarrow \lambda_i^* = 0 \end{cases}$
- **Karush-Kuhn-Tucker (KKT) conditions**:
  - *Primal constraints*: $f_i(x) \leq 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p$
  - *Dual constraints*: $\lambda \succeq 0$
  - *Complementary slackness*: $\lambda_i f_i(x) = 0, i = 1, \ldots, m$
  - Gradient of Lagrangian with respect to x is 0.
- For all problems:
  - x^*, λ^*, ν^* are optimal and strong duality holds → KKT conditions are satisfied.
- For convex problems:
  - x^*, λ^*, ν^* are optimal and strong duality holds ↔ KKT conditions are satisfied.
  - Slater's condition is satisfied → (x^* is optimal ↔ there exist λ^*, ν^* that satisfy KKT conditions).

2017-05-09

- Duality and problem reformulations:
  - Introduce new variables and equality constraints.
  - Make explicit constraints implicit or vice-versa.
  - Transform objective or constraint functions.
- Unconstrained problem: minimize $f_0(Ax + b)$
  - $g(\nu) = \begin{cases} -f_0^*(\nu) + b^\mathrm{T}\nu \text{ if } A^\mathrm{T}\nu = 0 \\ -\infty \text{ otherwise} \end{cases}$
  - Maximize $-f_0^*(\nu) + b^\mathrm{T}\nu$ subject to $A^\mathrm{T}\nu = 0$
- Norm approximation problem: minimize $\|Ax - b\|$
  - $g(\nu) = \begin{cases} b^\mathrm{T}\nu \text{ if } A^\mathrm{T}\nu = 0, \|\nu\|_* \leq 1 \\ -\infty \text{ otherwise} \end{cases}$
  - Maximize $b^\mathrm{T}\nu$ subject to $A^\mathrm{T}\nu = 0, \|\nu\|_* \leq 1$
- LP with box constraints: minimize $c^\mathrm{T}x$ subject to $Ax = b$ and $-1 \preceq x \preceq 1$
  - Reformulation: minimize $f_0(x) = \begin{cases} c^\mathrm{T}x \text{ if } -1 \preceq x \preceq 1 \\ \infty \text{ otherwise} \end{cases}$ subject to $Ax = b$.
  - $g(\nu) = \inf_{-1 \preceq x \preceq 1}(c^\mathrm{T}x + \nu^\mathrm{T}(Ax - b)) = -b^\mathrm{T}\nu - \|A^\mathrm{T}\nu + c\|_1$
  - Maximize $-b^\mathrm{T}\nu - \|A^\mathrm{T}\nu + c\|_1$
- Problems with generalized inequalities: minimize $f_0(x)$ subject to $f_i(x) \leq_{K_i} 0$ and $h_i(x) = 0$
  - Lagrange dual function, lower bound property, and Lagrange dual problem work in the same way.
- Semidefinite program: minimize $c^\mathrm{T}x$ subject to $\sum_{i=1}^n F_i x_i + G \preceq_K 0$ and $Ax = b$, where F and

G are semidefinite.

- Lagrange multiplier is a matrix $Z \in S^k$
- Lagrangian form $L(x, Z) = c^T x + \text{trace}(Z(x_1 F_1 + \cdots + x_n F_n - G))$
- $g(Z) = \inf_x L(x, Z) = \begin{cases} -\text{trace}(GZ) \text{ if } \text{trace}(F_i Z) + c_i = 0 \\ -\infty \text{ otherwise} \end{cases}$
- Maximize $-\text{trace}(GZ)$ subject to $Z \succeq 0$ and $\text{trace}(F_i Z) + c_i = 0$

- Maximum likelihood estimation: maximize $l(x) = \log p_x(y)$
  - x is a vector of unknown parameters to estimate.
  - y is a vector of observed value.
  - $l(x) = \log p_x(y)$ is called **log-likelihood function**.
- Linear measurements with IID noise:
  - $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$.
  - Maximize $l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$
  - Gaussian noise:
    - $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma 2)}$
    - Maximize $l(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (a_i^T x - y_i)^2$
    - x is the least-square solution
  - Laplacian noise:
    - $p(z) = \frac{1}{2a} e^{-|z|/a}$
    - Maximize $l(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^m |a_i^T x - y_i|$
    - x is the l1-norm solution
  - Uniform noise:
    - $p(z) = 1/(2a)$, where $z \in [-a, a]$
    - Maximize $l(x) = -m\log(2a)$ for all $|a_i^T x - y_i| \le a, i = 1, \ldots, m$
    - x is any solutions satisfying $|a_i^T x - y_i| \le a, i = 1, \ldots, m$
- **Logistic regression**:
  - $p(y) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$, where $y \in \{0, 1\}$
  - Maximzie $l(a, b) = \sum_{i=1}^k (a^T u_i + b) - \sum_{i=1}^m \log(1 + \exp(a^T u_i + b))$
  - a and b are unknown parameters to estimate from m observations u and y.
- [x] Homework: 5.1(a)-(c), 5.7(a)-(b), 5.27
- Solutions: Solutions4.pdf

2017-05-16

- Robust linear discrimination:

- Euclidean distance between hyperplanes $H_1 = \{z|a^{\mathrm{T}}z + b = 1\}$ and $H_2 = \{z|a^{\mathrm{T}}z + b = -1\}$ is $2/\|a\|_2$
- X_0 and X_1 are the data matrices classified as -1 and 1, respectively.
- Minimize $\|a\|_2/2$ subject to $X_0a + b \succeq 1$ and $X_1a + b \preceq -1$.
- Maximize $1^{\mathrm{T}}\lambda_0 + 1^{\mathrm{T}}\lambda_1$ subject to $2\|\lambda_0^{\mathrm{T}}X_0 - \lambda_1^{\mathrm{T}}X_1\|_2 \le 1$, $1^{\mathrm{T}}\lambda_0 = 1^{\mathrm{T}}\lambda_1$, $\lambda_0 \succeq 0$, $\lambda_1 \succeq 0$.
- Optimal value is distance between convex hulls.

- Approximate linear discrimination:
  - Minimze $1^{\mathrm{T}}\varepsilon_0 + 1^{\mathrm{T}}\varepsilon_1$ subject to $X_0a + b \succeq 1 - \varepsilon_0$ and $X_1a + b \preceq -1 + \varepsilon_1$, where $\varepsilon_0 \succeq 0, \varepsilon_1 \succeq 0$.
  - A heuristic for minimizing #misclassified points.
- Support vector classifier:
  - Minimze $\|a\|_2/2 + 1^{\mathrm{T}}\varepsilon_0 + 1^{\mathrm{T}}\varepsilon_1$ subject to $X_0a + b \succeq 1 - \varepsilon_0$ and $X_1a + b \preceq -1 + \varepsilon_1$, where $\varepsilon_0 \succeq 0, \varepsilon_1 \succeq 0$.
  - A trade-off curve between inverse of margin $2/\|a\|_2$ and classification error, measured by total slack $1^{\mathrm{T}}\varepsilon_0 + 1^{\mathrm{T}}\varepsilon_1$.
- Nonlinear discrimination:
  - Separate two sets of points by a nonlinear function: $f(x) = \theta^{\mathrm{T}}F(x)$ where $F(x)$ are a set of basis functions.
  - **Quadratic discrimination**: $f(z) = z^{\mathrm{T}}Pz + q^{\mathrm{T}}z + r$
  - **Polynomial discrimination**: $F(z)$ are all monomials up to a given degree.
- **Support vector machine (SVM)**:
  - Data matrix: $X = [\varphi(x_1)^{\mathrm{T}}, \ldots, \varphi(x_n)^{\mathrm{T}}], y \in \{-1, 1\}$, and φ is a discrimination function.
  - **Kernel tricks** $K(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j)$ avoids the explicit mapping from linear to nonlinear function. Let K denote XX^{T}.
  - Primal problem: minimize $\frac{1}{2}\omega^{\mathrm{T}}\omega$ subject to $\mathbf{diag}(y)(X\omega + b1) \succeq 1$
  - The Lagrangian: $L(\omega, b, \alpha) = \frac{1}{2}\omega^{\mathrm{T}}\omega - \alpha^{\mathrm{T}}(\mathbf{diag}(y)(X\omega + b1) - 1)$
  - Dual function: $\min_{\omega,b} L(\omega, b, \alpha) = \begin{cases} 1^{\mathrm{T}}\alpha - \frac{1}{2}\alpha^{\mathrm{T}}\mathbf{diag}(y)K\mathbf{diag}(y)\alpha \text{ if } y^{\mathrm{T}}\alpha = 0 \\ -\infty \text{ otherwise} \end{cases}$
  - Dual problem: maximize $1^{\mathrm{T}}\alpha - \frac{1}{2}\alpha^{\mathrm{T}}\mathbf{diag}(y)K\mathbf{diag}(y)\alpha$ subject to $y^{\mathrm{T}}\alpha = 0$ and $\alpha \succeq 0$
  - Decision function: $Xw + b = K\mathbf{diag}(y)\alpha + b = 0$
- Soft margin SVM:
  - Primal problem: minimize $\frac{1}{2}\omega^{\mathrm{T}}\omega + C1^{\mathrm{T}}\xi$ subject to $\mathbf{diag}(y)(X\omega + b1) \succeq 1 - \xi$ and $\xi \succeq 0$