Topics in Machine Learning Final

2017-06-06

- Logistic regression:
    - $p(y_i|x_i) = 1/(1+\exp(-y_i w^{\mathrm{T}} x_i))$, where $y_i \in \{-1, 1\}$
    - Log-likelihood function: $l(w) = -\sum_i \log(1+\exp(-y_i w^{\mathrm{T}} x_i))$
    - Primal problem: minimize $\sum_i \log(1+\exp(-y_i w^{\mathrm{T}} x_i))$
- Support vector machine:
    - Primal problem: minimize $\frac{1}{2}\omega^{\mathrm{T}}\omega + C\sum_i \xi_i$ subject to $y_i w^{\mathrm{T}} x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$
- Loss functions $\xi(w; x, y)$:
    - Ideal: $\xi(w; x_i, y_i) = \begin{cases} 1 \text{ if } y_i w^{\mathrm{T}} x_i < 0 \\ 0 \text{ otherwise} \end{cases}$
    - Logistic regression: $\xi(w; x_i, y_i) = \log(1+\exp(-y_i w^{\mathrm{T}} x_i))$
    - Support vector machine: $\xi(w; x_i, y_i) = \max\{0, 1 - y_i w^{\mathrm{T}} x_i\}$
- Regularization $R(w)$:
    - To avoid overfitting by manipulating the value of w to be less extreme.
    - The idea of making w closer to zero is to create the maximal margin in SVM.
    - C is a regularization parameter arbitrarily chosen by users to weigh the importance of the regularization term.
    - Support vector machine: $R(w) = \frac{1}{2}\omega^{\mathrm{T}}\omega$, which is an L2 regularization.

2017-06-13

- Unconstrained minimization assumes
    - f is convex and twice continuously differentiable.
    - Optimal value $p^* = \inf_x f(x)$ is attained.
- Descent methods require a starting point $x^{(0)}$ such that
    - $x^{(0)} \in \mathbf{dom} f$.
    - Sublevel set $S = \{x | f(x) \leq f(x^{(0)})\}$ is closed
        - S is closed if $\mathbf{dom} f = \mathbb{R}^n$.
        - S is closed if $f(x) \to \infty$ as $x \to$ boundary of $\mathbf{dom} f$.
- f is **strongly convex** on S:
    - There exists an $m > 0$ such that $\nabla^2 f(x) \succeq mI$ for all $x \in S$.

- $f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \frac{m}{2}\|y-x\|_2^2$ for $x, y \in S$.
  - Stopping criterion: $f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2$.
- **Descent methods**:
  - $x^+ = x + t\Delta x$ with $f(x^+) < f(x)$ at each step, where $t > 0$ is the step size.
  - From convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^{\mathrm{T}}\Delta x < 0$ (i.e., Δx is a *descent direction*).
  - Stopping criterion with parameter ε: $\|\nabla f(x)\|_2 \leq \varepsilon$ implies $f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2 \leq \frac{\varepsilon^2}{2m}$
- Line search:
  - **Exact line search**: $t = \mathbf{argmin}_{t>0} f(x + t\Delta x)$
  - **Backtracking line search** with parameters $\alpha \in (0, 1/2), \beta \in (0, 1)$: starting at $t = 1$, repeat $t := \beta t$ until $f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^{\mathrm{T}}\Delta x$
- **Gradient descent method**:
  - $\Delta x = -\nabla f(x)$
  - Convergence with exact line search: if $MI \succeq \nabla^2 f(x) \succeq mI$, then $f(x^+) - p^* \leq (1 - \frac{m}{M})(f(x) - p^*)$ or $f(x^{(k)}) - p^* \leq (1 - \frac{m}{M})^k(f(x^{(0)}) - p^*)$
- **Steepest descent method**:
  - Normalized steepest descent direction: $\Delta x_{\mathrm{nsd}} = \mathbf{argmin}\{\nabla f(x)^{\mathrm{T}}v \mid \|v\| = 1\}$
  - (Unnormalized) steepest descent direction: $\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_* \Delta x_{\mathrm{nsd}}$
  - Euclidean norm: $\Delta x_{\mathrm{sd}} = -\nabla f(x)$
  - Quadratic norm: $\Delta x_{\mathrm{sd}} = -P^{-1}\nabla f(x)$, where $\|x\|_P = (x^{\mathrm{T}}Px)^{1/2}, P \in \mathbb{S}_{++}^n$
  - l1-norm: $\Delta x_{\mathrm{sd}} = -\frac{\partial f(x)}{\partial x_i}e_i$, where $|\frac{\partial f(x)}{\partial x_i}| = \|\nabla f(x)\|_\infty$
- **Newton step**:
  - $\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1}\nabla f(x)$
  - $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation $f(x + v) = f(x) + \nabla f(x)^{\mathrm{T}}v + \frac{1}{2}v^{\mathrm{T}}\nabla^2 f(x)v$
  - $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition $\nabla f(x + v) \approx \nabla \hat{f}(x + v) = \nabla f(x) + \nabla^2 f(x)v = 0$
  - Δx_{nt} is the steepest descent direction at x in local Hessian norm $\|u\|_{\nabla^2 f(x)} = (u^{\mathrm{T}}\nabla^2 f(x)u)^{1/2}$