CAP 6610 Machine Learning, Spring 2020

Homework 3

Due 3/20/2020 11:59PM

1. *MAP interpretation of regularized empirical loss minimization.* We have seen that some (unregularized) empirical loss minimization problems can be interpreted as maximum likelihood estimation (MLE) if we choose certain parametric form for the conditional probability $p(y|\boldsymbol{x};\boldsymbol{\theta})$. Assuming the data samples are i.i.d., MLE of $p(y|\boldsymbol{x};\boldsymbol{\theta})$ is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{n} -\log p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}).$$

After some trivial transformations, we can recover some supervised learning models such as least squares regression and logistic classification.

Some statisticians, who call themselves Bayesians, believe that we should treat $\boldsymbol{\theta}$ as random as well, and impose probability distributions on them. In this case, the probability that we really care about is $p(\boldsymbol{\theta}|Y,\boldsymbol{X})$, the conditional probability of $\boldsymbol{\theta}$ given data $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ and $Y=\{y_1,\ldots,y_n\}$. According to Bayes rule,

$$p(\boldsymbol{\theta}|Y,\boldsymbol{X}) = \frac{p(Y|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})}{p(Y|\boldsymbol{X})}.$$

Furthermore, it is common to assume that $\boldsymbol{\theta}$ is independent of $\boldsymbol{X}$ and $(\boldsymbol{x}_i,y_i)$ are i.i.d. conditioned on $\boldsymbol{\theta}$, leading to

$$p(\boldsymbol{\theta}|Y,\boldsymbol{X}) = \frac{p(\boldsymbol{\theta})\prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})}{p(Y|\boldsymbol{X})}.$$

Here, $p(\boldsymbol{\theta})$ is called the prior (*a priori* in Latin), $p(y|\boldsymbol{x},\boldsymbol{\theta})$ is called the likelihood, and $p(\boldsymbol{\theta}|Y,\boldsymbol{X})$ is called the posterior (*a posteriori* in Latin).

Depending on the definition of the prior and the likelihood, the denominator $p(Y|\boldsymbol{X})$ may be very hard to evaluate. Instead, we can try to find a point estimate $\boldsymbol{\theta}$ that maximizes the posterior probability, which is called maximum *a posteriori* (MAP), since the denominator does not depend on $\boldsymbol{\theta}$ and can be omitted in maximization. This is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{n} -\log p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}).$$

For each of the following cases, given an explicit MAP formulation for estimating $\boldsymbol{\theta}$. Find their relationship to the corresponding regularized empirical loss minimization problems. Specifically, give an exact expression for the regularization parameter λ in terms of the prior and likelihood distributions.

(a) $p(y|\boldsymbol{x},\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\phi}^\top\boldsymbol{\theta}, \sigma^2)$ and $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, \sigma_0^2\boldsymbol{I})$;

(b) $p(y|\boldsymbol{x},\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\phi}^\top\boldsymbol{\theta}, \sigma^2)$ and $p(\boldsymbol{\theta})$ follows a multivariate Laplacian distribution:

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{m} \frac{1}{2a} \exp\left(-\frac{|\theta_j|}{a}\right);$$

(c) $p(y|\boldsymbol{x},\boldsymbol{\theta}) = \Pr[yu \geq 0]$ where $y = \pm 1$, $p(u|\boldsymbol{x},\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\phi}^\top\boldsymbol{\theta}, \sigma^2)$ and $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, \sigma_0^2\boldsymbol{I})$;

(d) $p(y|\boldsymbol{x},\boldsymbol{\theta}) = 1/(1+\exp(-y\boldsymbol{\phi}^\top\boldsymbol{\theta}))$ where $y = \pm 1$ and $p(\boldsymbol{\theta})$ follows a multivariate Laplacian distribution as in (b).

2. *Nonexpansiveness of proximal operators.* In this problem we show that for a convex function f (not necessarily differentiable), its proximal operator is nonexpansive, i.e.,

$$\|\operatorname{Prox}_f(\boldsymbol{\theta}_1) - \operatorname{Prox}_f(\boldsymbol{\theta}_2)\| \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

where

$$\operatorname{Prox}_f(\boldsymbol{\theta}_1) = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|^2,$$

with the following steps:

(a) Show that

$$\boldsymbol{\theta}_1 - \operatorname{Prox}_f(\boldsymbol{\theta}_1) \in \partial f(\operatorname{Prox}_f(\boldsymbol{\theta}_1)).$$

(b) Show that if $\boldsymbol{g}_1 \in \partial f(\boldsymbol{\theta}_1)$ and $\boldsymbol{g}_2 \in \partial f(\boldsymbol{\theta}_2)$, then

$$(\boldsymbol{g}_1 - \boldsymbol{g}_2)^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq 0.$$

Hint. By definition, if $\boldsymbol{g}_1$ is a subgradient of $f(\boldsymbol{\theta}_1)$, then for all $\boldsymbol{\theta}$

$$f(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}_1) + \boldsymbol{g}_1^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_1).$$

(c) Use the previous two results to show the *firm nonexpansiveness*

$$(\operatorname{Prox}_f(\boldsymbol{\theta}_1) - \operatorname{Prox}_f(\boldsymbol{\theta}_2))^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq \|\operatorname{Prox}_f(\boldsymbol{\theta}_1) - \operatorname{Prox}_f(\boldsymbol{\theta}_2)\|^2.$$

(d) Apply the Cauchy-Schwartz inequality to obtain the nonexpansiveness property.

3. *Hand-written digits classification.* The MNIST data set is a famous data set for multi-class classification, which can be downloaded here `http://yann.lecun.com/exdb/mnist/`. In this problem you will design a SGD algorithm for multi-class support vector machine with group-sparse regularization that solves the following optimization problem

$$\underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} \max_c(\boldsymbol{x}_i^\top\boldsymbol{\theta}_c - \boldsymbol{x}_i^\top\boldsymbol{\theta}_{y_i} + 1_{y_i \neq c}) + \lambda \sum_{j=1}^{m} \sqrt{\sum_{c=1}^{k} \theta_{jc}^2}.$$

Here we simply assume that the features are the image pixels themselves (we even ignore the constant 1 here).

(a) Derive the stochastic proximal subgradient algorithm for solving it. For simplicity, you can assume that there is only one term that reaches the maximum value in $\max_c(\boldsymbol{\phi}^\top \boldsymbol{\theta}_c - \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{y_i} + 1_{y_i \neq c})$ throughout the iterations. At iteration t, you can simply denote the step size as $\gamma^{(t)}$.

(b) Implement the algorithm in your favorite programming language.

(c) Run the algorithm with $\lambda = 10, 1, 0.1, 0.01$ and diminishing step size $\gamma^{(t)} = 1/t$, and run the algorithm for 10^6 iterations. At every 1000 iteration, evaluate the prediction accuracy on the test set and plot the progress on a figure.

(d) For the solution of each λ value, show a black and white figure for the pixels that are being used to make the predictions. Is it true that a large λ leads to a more sparse solution?