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).

Solution. We fix the regularized empirical loss minimization formulation to be

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta}). \tag{1}$$

Sometimes the $1/n$ factor is dropped, but that basically corresponds to a regularization parameter $n\lambda$. Either way is acceptable.

(a) The log-likelihood of $p(y|\boldsymbol{x},\boldsymbol{\theta})$ is

$$\log p(y|\boldsymbol{x},\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\boldsymbol{\phi}^\top\boldsymbol{\theta})^2.$$

The log-likelihood of $p(\boldsymbol{\theta})$ is

$$\log p(\boldsymbol{\theta}) = -\frac{m}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\|\boldsymbol{\theta}\|^2.$$

Therefore, the MAP formulation is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\phi}_i^\top\boldsymbol{\theta})^2 + \frac{\sigma^2}{n\sigma_0^2}\|\boldsymbol{\theta}\|^2.$$

This is ridge regression with $\lambda = \sigma^2/(n\sigma_0^2)$.

(b) The log-likelihood of $p(y|\boldsymbol{x},\boldsymbol{\theta})$ is the same as part (a). The log-likelihood of $p(\boldsymbol{\theta})$ is

$$\log p(\boldsymbol{\theta}) = -m\log(2a) - \frac{1}{a}\|\boldsymbol{\theta}\|_1.$$

Therefore, the MAP formulation is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\phi}_i^\top\boldsymbol{\theta})^2 + \frac{\sigma^2}{na}\|\boldsymbol{\theta}\|_1.$$

This is lasso regression with $\lambda = \sigma^2/(na)$.

(c) The log-likelihood of $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ is

$$\begin{aligned}\log p(y|\boldsymbol{x}, \boldsymbol{\theta}) &= \log \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(u - y\boldsymbol{\phi}^\top \boldsymbol{\theta})^2}{2}\right) du \\ &= \log \int_{-\infty}^{y\boldsymbol{\phi}^\top \boldsymbol{\theta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \\ &= \log \Phi(y\boldsymbol{\phi}^\top \boldsymbol{\theta}),\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of $\mathcal{N}(0, 1)$. The log-likelihood of $p(\boldsymbol{\theta})$ is the same as in part (a). The MAP formulation is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} -\log \Phi(y_i \boldsymbol{\phi}_i^\top \boldsymbol{\theta}) + \frac{1}{n\sigma_0^2} \|\boldsymbol{\theta}\|^2.$$

This is L_2 regularized probit regression with $\lambda = 1/(n\sigma_0^2)$.

(d) The log-likelihood of $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ is

$$\log p(y|\boldsymbol{x}, \boldsymbol{\theta}) = -\log(1 + \exp(-y\boldsymbol{\phi}^\top \boldsymbol{\theta})).$$

The log-likelihood of $p(\boldsymbol{\theta})$ is the same as in part (b). The MAP formulation is equivalent to

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\phi}_i^\top \boldsymbol{\theta})) + \frac{1}{na} \|\boldsymbol{\theta}\|^2.$$

This is L_1 regularized logistic regression with $\lambda = 1/(na)$.

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*

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

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

Solution.

(a) By definition, $\mathrm{Prox}_f(\boldsymbol{\theta}_1)$ minimizes $f(\boldsymbol{\theta}) + (1/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|^2$, so the subgradient optimality condition implies

$$0 \in \partial\left(f(\mathrm{Prox}_f(\boldsymbol{\theta}_1)) + \frac{1}{2}\|\mathrm{Prox}_f(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_1\|^2\right).$$

The right-hand-side is a set defined as

$$\{\boldsymbol{g} + \mathrm{Prox}_f(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_1 \mid \boldsymbol{g} \in \partial f(\mathrm{Prox}_f(\boldsymbol{\theta}_1))\}.$$

This means there exists a vector $\boldsymbol{g} \in \partial f(\mathrm{Prox}_f(\boldsymbol{\theta}_1))$ such that

$$\boldsymbol{g} + \mathrm{Prox}_f(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_1 = 0,$$

which is equivalent to

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

(b) Apply the hint with $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ we get

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

Similarly, if $\boldsymbol{g}_2 \in \partial f(\boldsymbol{\theta}_2)$, then

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

Add up these two inequalities gives

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

(c) Combining (a) and (b) gives us

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

Slightly rearranging it gives us

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

(d) Cauchy-Schwartz implies

$$(\mathrm{Prox}_f(\boldsymbol{\theta}_1) - \mathrm{Prox}_f(\boldsymbol{\theta}_2))^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \leq \|\mathrm{Prox}_f(\boldsymbol{\theta}_1) - \mathrm{Prox}_f(\boldsymbol{\theta}_2)\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Combining it with (c) we have

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

Divide both sides with $\|\mathrm{Prox}_f(\boldsymbol{\theta}_1) - \mathrm{Prox}_f(\boldsymbol{\theta}_2)\| > 0$ shows 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?

Solution.

(a) Denote $\boldsymbol{\Theta} = [\ \boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \cdots\ \boldsymbol{\theta}_k\ ]$ and

$$\ell_i(\boldsymbol{\Theta}) = \max_{c}(\boldsymbol{x}_i^\top\boldsymbol{\theta}_c - \boldsymbol{x}_i^\top\boldsymbol{\theta}_{y_i} + 1_{y_i\neq c}).$$

Suppose at a particular point $\boldsymbol{\Theta}^{(t)}$ only one of them attains the maximum, then the function is differentiable. If maximum is attained at $y_i = \arg\max_c$, then

$$\nabla\ell_i(\boldsymbol{\Theta}^{(t)}) = 0;$$

otherwise, denote $\widehat{y_i} = \arg\max_c$, and

$$\nabla_{\boldsymbol{\theta}_c}\ell_i(\boldsymbol{\Theta}^{(t)}) = \begin{cases}\boldsymbol{x}_i & c = \widehat{y_i}\\ -\boldsymbol{x}_i & c = y_i\\ 0 & \text{otherwise.}\end{cases}$$

Intuitively, the gradient descent step updates at most two columns of the $\boldsymbol{\Theta}$ matrix: the column that corresponds to the correct label is added with a little bit of $\boldsymbol{x}_i$, and the

column corresponding to the predicted label is subtracted with a little bit of $\boldsymbol{x}_i$; all the other columns stay the same.

Then we apply row-wise block soft-thresholding on $\boldsymbol{\Theta}$. The norm of each row of $\boldsymbol{\Theta}$ is calculated; if it is less than $\gamma^{(t)}\lambda$, the entire row is set to zero, otherwise the magnitude of the row decreases by the amount of $\gamma^{(t)}\lambda$.

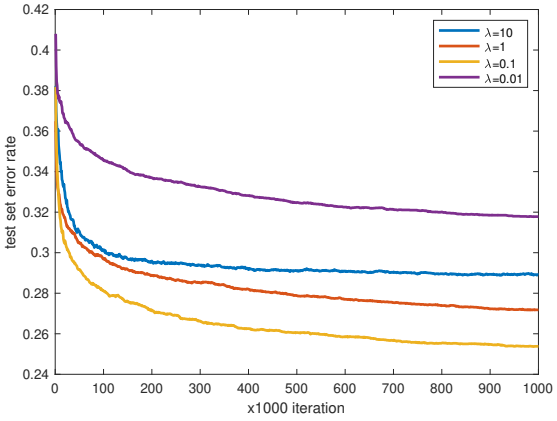
The stochastic proximal subgradient algorithm for L_1-norm regularized multiclass SVM is elaborated here. The cth column of $\boldsymbol{\Theta}$ is denoted $\boldsymbol{\theta}_c$, and the jth row is denoted $\boldsymbol{\Theta}_j$.

---

initialize $\boldsymbol{\Theta}^{(0)}$
for $t = 0, 1, \ldots$ **do**
  uniformly sample i from $\{1, \ldots, n\}$
  calculate $\widehat{y}_i = \arg\max_c (\boldsymbol{x}_i^\top \boldsymbol{\theta}_c^{(t)} - \boldsymbol{x}_i^\top \boldsymbol{\theta}_{y_i}^{(t)} + 1_{y_i \neq c})$
  $\boldsymbol{\theta}_{y_i}^{(t+1)} \leftarrow \boldsymbol{\theta}_{y_i}^{(t)} + \gamma^{(t)} \boldsymbol{x}_i$
  $\boldsymbol{\theta}_{\widehat{y}_i}^{(t+1)} \leftarrow \boldsymbol{\theta}_{\widehat{y}_i}^{(t)} - \gamma^{(t)} \boldsymbol{x}_i$
  for $j = 1$ **to** m **do**
    $\nu \leftarrow \|\boldsymbol{\Theta}_j^{(t+1)}\|$
    if $\nu \leq \gamma^{(t)}\lambda$ **then**
      $\boldsymbol{\Theta}_j^{(t+1)} \leftarrow 0$
    else
      $\boldsymbol{\Theta}_j^{(t+1)} \leftarrow \left(1 - \dfrac{\gamma^{(t)}\lambda}{\nu}\right) \boldsymbol{\Theta}_j^{(t+1)}$
    end if
  end for
end for

(b) A simple MATLAB code is given.

(c) Here is the convergence plot for four different λ values using the same initialization.



The testing error rate goes down as the algorithm progresses, which is a good sign that the method works. Moreover, we do see the pattern that as we reduce the value of λ, the testing error rate decreases at first and then goes back up.

(d) For each λ we obtain a $\boldsymbol{\Theta}$ solution, and each row of $\boldsymbol{\Theta}$ corresponds to one pixel. In the following figure, each pixel is either black or white, indicating whether the corresponding row in $\boldsymbol{\Theta}$ is zero (white) or not (black). Indeed, we see that a larger λ leads to fewer pixels being used for making the prediction; in fact for $\lambda = 0.1$ and $\lambda = 0.01$ all pixels are being used. When we do select pixels to be ignored for making the predictions, they all lie outside the center circle, which do make sense since we seldom write outside that region.