CAP 6610 Machine Learning, Spring 2025

Homework 3

Due Mar. 25, 11:59pm

1. What is the distance between two parallel hyperplanes $\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{w}^\top \boldsymbol{x} = \beta_1\}$ and $\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{w}^\top \boldsymbol{x} = \beta_2\}$? *Hint.* Let $\boldsymbol{w}^\top \boldsymbol{x}_1 = \beta_1$, $\boldsymbol{w}^\top \boldsymbol{x}_2 = \beta_2$, and minimize $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2$.

2. *Log-concavity of Gaussian cumulative distribution function.* The cumulative distributionfunction of a Gaussian random variable,

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$$

is log-concave. This follows from the general result that the convolution of two log-concave functions is log-concave. In this problem we guide you through a simple self-contained proof that Φ is log-concave.

We will use the fact that Φ is log-concave if and only if $\Phi^{''}(t)\Phi(t) \leq (\Phi^{'}(t))^2$. This comes from the definition that $(\log \Phi(t))^{''} \leq 0$ everywhere if Φ is log-concave. Applying the chain rule, we have

$$(\log \Phi(t))^{'} = \frac{\Phi^{'}(t)}{\Phi(t)}.$$

Its second derivative is

$$(\log \Phi(t))^{''} = \frac{\Phi^{''}(t)}{\Phi(t)} - \frac{(\Phi^{'}(t))^2}{(\Phi(t))^2}.$$

Letting it nonpositive gives the condition $\Phi^{''}(t)\Phi(t) \leq (\Phi^{'}(t))^2$.

   (a) Verify that $\Phi^{''}(t)\Phi(t) \leq (\Phi^{'}(t))^2$ for $t \geq 0$. That leaves us the hard part, which is to show the inequality for $t < 0$.

   (b) Verify that for any t and x we have $x^2/2 \geq -t^2/2 + tx$.

   (c) Using part (b) to show that $e^{-x^2/2} \leq e^{t^2/2 - tx}$. Conclude that

$$\int_{-\infty}^{t} e^{-x^2/2} dx \leq e^{t^2/2} \int_{-\infty}^{t} e^{-tx} dx.$$

   (d) Use part (c) to verify that $\Phi^{''}(t)\Phi(t) \leq (\Phi^{'}(t))^2$ for $t < 0$.

3. *Maximum likelihood estimation for exponential family.* A probability distribution or density on a set $\mathcal{X}$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^m$, is called an *exponential family* if it has the form

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = a(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})),$$

for $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^m$, and $a(\boldsymbol{\theta})$ is a normalization function. Here we interpret as a density function when $\mathcal{X}$ is a continuous set, and a probability distribution if $\mathcal{X}$ is discrete. Thus we have

$$a(\boldsymbol{\theta}) = \left(\int_{\mathcal{X}} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})) d\boldsymbol{x} \right)^{-1}$$

when $p(\boldsymbol{x};\boldsymbol{\theta})$ is a density, and

$$a(\boldsymbol{\theta}) = \left(\sum_{\boldsymbol{x}\in\mathcal{X}} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}))\right)^{-1}$$

when $p(\boldsymbol{x};\boldsymbol{\theta})$ represents a distribution. We consider only values of $\boldsymbol{\theta}$ for which the integral or sum above is finite. Many families of distributions have this form, for appropriate choice of the parameter $\boldsymbol{\theta}$ and function $\boldsymbol{\phi}$.

Show that for any $\boldsymbol{x} \in \mathcal{X}$, the log-likelihood function $\log p(\boldsymbol{x};\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$. This means that maximum-likelihood estimation for an exponential family leads to a convex optimization problem. You don't have to give a formal proof of concavity of $\log p(\boldsymbol{x};\boldsymbol{\theta})$ in the general case: You can just consider the case when $\mathcal{X}$ is finite, and state that the other cases (discrete but infinite $\mathcal{X}$, continuous $\mathcal{X}$) can be handled by taking limits of finite sums.

4. We test the performance of three regression methods on the wine data set `http://archive.ics.uci.edu/ml/datasets/Wine+Quality` that we have seen in Homework 1, Question 4. All models are trained by solving the following optimization problem

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \sum_{i=1}^{n} \ell(\boldsymbol{x}_i^\top \boldsymbol{w} + \beta - y_i),$$

where the loss functions are

- least squares loss $\ell(t) = t^2$
- Huber loss
$$\ell(t) = \begin{cases} t^2 & |t| \le 1/2 \\ |t| - 1/4 & |\epsilon| > 1/2 \end{cases}$$
- hinge loss
$$\ell(t) = \begin{cases} 0 & |t| \le 1/2 \\ |t| - 1/2 & |t| > 1/2 \end{cases}$$

The least squares loss can be directly solved by the command `Phi\y` for some properly defined `Phi`. For the latter two, you will use the `cvx` package found on Prof. Boyd's website `https://web.stanford.edu/~boyd/software.html`. Report their prediction performance on the test set using a different metric, mean absolution error (MAE), defined as $(1/n)\sum_{i=1}^{n} |y_i - \hat{y}_i|$.

Hint. For the Huber loss and hinge loss, consider rewriting them as $\max(\cdot)$ of some convex and smooth functions. `cvx` is able to recognize them as valid convex functions and carry out the computations.

5. We test the performance of three classification methods on the ionosphere data set `https://archive.ics.uci.edu/ml/datasets/ionosphere` that we have seen in Homework 1, Question 5. All models are trained by solving the following optimization problem

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \sum_{i=1}^{n} \ell(\boldsymbol{x}_i^\top \boldsymbol{w} + \beta, y_i),$$

where the loss functions are

- least squares loss $\ell(t, y) = (yt - 1)^2$
- logistic loss $\ell(t, y) = \log(1 + \exp(-yt))$
- hinge loss $\ell(t, y) = \max(0, 1 - yt)$

Again, you will use the backslash command to solve for the first model, and `cvx` to solve for the latter two. Report their prediction accuracy on the test set.