Projects
Machine Learning Course
Fall 2016

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
mlo.epfl.ch/page-136795.html
epfmlcourse@gmail.com

Mock Midterm Exam - Nov 15, 2016

1 Subgradient Descent [20pts]

Derive the (sub)gradient descent update rule for a one-parameter linear model using the Mean Absolute Error,

$$\mathcal{L}_{\mathsf{MAE}}(\mathbf{X}, \mathbf{y}, w) = \frac{1}{N} \sum_{n=1}^{N} |wx_n - y_n|.$$

Hint: The function $f(x) = |ax|$ is a composition of two simpler function. Use the chain rule!

2 Multiple-Output Regression [20pts]

Let $S = \{(\mathbf{y}_n, \mathbf{x}_n)\}_{n=1}^N$ be our training set for a regression problem with $\mathbf{x}_n \in \mathbb{R}^D$ as usual. But now $\mathbf{y}_n \in \mathbb{R}^K$, i.e., we have K outputs for each input. We want to fit a linear model for each of the K outputs, i.e., we now have K regressors $f_k(\cdot)$ of the form

$$f_k(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}_k,$$

where $\mathbf{w}_k^\top = (\mathbf{w}_{k1}, \cdots, \mathbf{w}_{kD})$ is the weight vector corresponding to the k-th regressor. Let $\mathbf{W}$ be the $D \times K$ matrix whose columns are the vectors $\mathbf{w}_k$.

Our goal is to minimize the following cost function $\mathcal{L}$:

$$\mathcal{L}(\mathbf{W}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{1}{2\sigma_k^2} (y_{nk} - \mathbf{x}_n^\top \mathbf{w}_k)^2 + \frac{1}{2} \sum_{k=1}^{K} \|\mathbf{w}_k\|_2^2,$$

where the σ_k are known real-valued scalars. Let $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_K)$.

For the solution, let $\mathbf{X}$ be the $D \times N$ matrix whose columns are the feature vectors $\mathbf{x}_n$.

1. (4pt) Write down the normal equations for $\mathbf{W}^\star$, the minimizer of the cost function. I.e., what is the first-order condition that $\mathbf{W}^\star$ has to fulfill in order to minimize $\mathcal{L}(\mathbf{W})$.
2. (8pts) Is the minimum $\mathbf{W}^\star$ unique? Assuming it is, write down an expression for this unique solution.
3. (8pts) Write down a probabilistic model, so that the MAP solution for this model coincides with minimizing the above cost function. Note that this will involve specifying the the likelihoods as well as a suitable prior (which will give you the regression term).

3 Proportional Hazard Model [20pts]

Let $S = \{(y_n, \mathbf{x}_n)\}_{n=1}^N$ be our training set for a regression problem with $\mathbf{x}_n \in \mathbb{R}^D$ as usual. We assume that the output y_n is *ordered*, i.e., takes values in the set $\{1, 2, \dots, K\}$ where we think of these numbers as *ordered* by the natural ordering. We wish to fit a linear model.

In the *proportional hazard* model we use the following probability distribution,

$$p(y_n = k \mid \mathbf{x}_n, \mathbf{w}, \mathbf{\Theta}) = \frac{e^{\eta_{nk}}}{\sum_{j=1}^K e^{\eta_{nj}}},$$

where $\eta_{nk} = \Theta_k + \mathbf{x}_n^\top \mathbf{w}$. The scalars Θ_k are assumed to be ordered, i.e., $\Theta_1 > \Theta_2 \cdots > \Theta_K$. Let $\mathbf{\Theta} = (\Theta_1, \cdots, \Theta_K)$.

1. (4pts) Show that $p(y_n \,|\, \mathbf{x}_n, \mathbf{w}, \mathbf{\Theta})$ (and therefore also $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \mathbf{\Theta})$) is a valid distribution.

   Hint: What are the *two* conditions that you need to verify?

2. (8pts) Derive the log-likelihood for this model.

3. (8pts) Show that the negative of the log-likelihood is convex with respect to $\mathbf{\Theta}$ and $\mathbf{w}$.

   HINT: You can assume that the function $\ln(\sum_{k=1}^K e^{t_k})$ is convex.

4 Multiple Choice Questions and Simple Problems [40pts]

Mark the correct **answer(s)**. More than one answer can be correct!

- In regression, "complex" models tend to
  1. (1 pt) overfit
  2. (1 pt) have large bias
  3. (1 pt) have large variance
- In regression, "simple" models tend to
  1. (1 pt) overfit
  2. (1 pt) have large bias
  3. (1 pt) have large variance
- We sometimes add a regularization term because
  1. (1 pt) this sometimes renders the minimization problem of the cost function into a strictly convex/concave problem
  2. (1 pt) this tends to avoid overfitting
  3. (1 pt) this converts a regression problem into a classification problem
- The k-nearest neighbor classifier
  1. (1 pt) typically works the better the larger the dimension of the feature space
  2. (1 pt) can classify up to k classes
  3. (1 pt) typically works the worse the larger the dimension of the feature space
  4. (1 pt) can only be applied if the data can be linearly separated
  5. (1 pt) has a misclassification rate of at most two times the one of the Bayes classifier if we have lots of data
  6. (1 pt) has a misclassification rate that is two times better than the one of the Bayes classifier
- A real-valued scalar Gaussian distribution
  1. (1 pt) is a member of the exponential family with one scalar parameter
  2. (1 pt) is a member of the exponential family with two scalar parameters
  3. (1 pt) is not a member of the exponential family
- Which of the following statements is correct, where we assume that all the stated minima and maxima are in fact taken on in the domain of relevance.
  1. (1 pt) $\max\{0, x\} = \max_{\alpha \in [0,1]} \alpha x$
  2. (1 pt) $\min\{0, x\} = \min_{\alpha \in [0,1]} \alpha x$
  3. (1 pt) Let $g(x) := \min_y f(x, y)$. Then $g(x) \leq f(x, y)$
  4. (1 pt) $\max_x g(x) \leq \max_x f(x, y)$
  5. (1 pt) $\max_x \min_y f(x, y) \leq \min_y \max_x f(x, y)$
- Which of the following statements are correct?
  1. (1 pt) The training error is typically smaller than the test error.
  2. (1 pt) The original SVM formulation can be optimized using SGD.
  3. (1 pt) One iteration of SGD for ridge regression costs roughly $\Theta(ND)$.
  4. (1 pt) The original logistic regression formulation can be optimized using SGD.
  5. (1 pt) We discussed coordinate descent to optimize the original SVM formulation.

- The following functions are convex:
  1. (1 pt) $f(x) := x^2$, $x \in \mathbb{R}$
  2. (1 pt) $f(x) := x^3$, $x \in [-1, 1]$
  3. (1 pt) $f(x) := -x^3$, $x \in [-1, 0]$
  4. (1 pt) $f(x) := e^{-x}$, $x \in \mathbb{R}$
  5. (1 pt) $f(x) := e^{-x^2/2}$, $x \in \mathbb{R}$
  6. (1 pt) $f(x) := \ln(1/x)$, $x \in [0, \infty)$
  7. (1 pt) $f(x) := g(h(x))$, $x \in \mathbb{R}$, where g, h are convex and increasing over $\mathbb{R}$
- (5 pts) Let $f : \mathbb{R}^D \to \mathbb{R}$ be the function $f(\mathbf{w}) := \exp(\mathbf{x}^\top \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$. What is $\nabla_{\mathbf{w}} f$?