10-401 Machine Learning: Homework 3

Due 5:00 p.m. Monday, March 5, 2018

Instructions

- **Submit your homework on time electronically by submitting to Autolab by 5:00 pm, Monday, March 5, 2018**.

  We recommend that you use LaTeX, but we will accept scanned solutions as well. This assignment contains no programming questions, so you do not need to download any extra files from Autolab. To submit this homework, you should submit a pdf of your solutions on Autolab by navigating to Homework 3 and clicking the "Submit File" button.

- **Late homework policy**: Homework is worth full credit if submitted before the due date. Up to 50 % credit can be received if the submission is less than 48 hours late. The lowest homework grade at the end of the semester will be dropped. Please talk to the instructor in the case of extreme extenuating circumstances.

- **Collaboration policy**: You are welcome to collaborate on any of the questions with anybody you like. However, you must write up your own final solution, and you must list the names of anybody you collaborated with on this assignment.

Problem 1: Logistic Regression and Gradient Descent

1.1: Logistic regression in two dimensions

In this question, we will derive the logistic regression algorithm (the M(C)LE and its gradient). For simplicity, we assume the dataset is two-dimensional. Given a training set $\{(x^i, y^i), i = 1, \ldots, n\}$ where $x^i \in \mathbb{R}^2$ is a feature vector and $y^i \in \{0, 1\}$ is a binary label, we want to find the parameters $\hat{w}$ that maximize the likelihood for the training set, assuming a parametric model of the form

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)} = \frac{\exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)}. \tag{1}$$

1. **[10 Points]** Below, we give a derivation of the conditional log likelihood. In this derivation, **provide a short justification for why each line follows from the previous one**.

$$\begin{aligned}
\ell(w) &\equiv \ln \prod_{j=1}^{n} p(y^j \mid x^j, w) && (2)\\
&= \sum_{j=1}^{n} \ln p(y^j \mid x^j, w) && (3)\\
&= \sum_{j=1}^{n} \ln \left(p(y^j = 1 \mid x^j, w)^{y^j} p(y^j = 0 \mid x^j, w)^{1-y^j} \right) && (4)\\
&= \sum_{j=1}^{n} \left[y^j \ln p(y^j = 1 \mid x^j, w) + (1 - y^j) \ln p(y^j = 0 \mid x^j, w) \right] && (5)\\
&= \sum_{j=1}^{n} \left[y^j \ln \frac{\exp(w_0 + w_1 x_1^j + w_2 x_2^j)}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} + (1 - y^j) \ln \frac{1}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} \right] && (6)\\
&= \sum_{j=1}^{n} \left[y^j \ln \left(\exp(w_0 + w_1 x_1^j + w_2 x_2^j) \right) + \ln \left(\frac{1}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} \right) \right] && (7)\\
&= \sum_{j=1}^{n} \left[y^j \left(w_0 + w_1 x_1^j + w_2 x_2^j \right) - \ln \left(1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j) \right) \right]. && (8)
\end{aligned}$$

Next, we will derive the gradient of the previous expression with respect to w_0, w_1, w_2, i.e., $\frac{\partial \ell(w)}{\partial w_i}$, where $\ell(w)$ denotes the log likelihood from part 1. We will perform a few steps of the derivation, and then ask you to do one step at the end. If we take the derivative of Expression 8 with respect to w_i for $i \in \{1, 2\}$, we get the following expression:

$$\frac{\partial \ell(w)}{\partial w_i} = \frac{\partial}{\partial w_i} \sum_{j=1}^{n} \left[y^j \left(w_0 + w_1 x_1^j + w_2 x_2^j \right) \right] - \frac{\partial}{\partial w_i} \sum_{j=1}^{n} \ln \left[1 + \exp \left(w_0 + w_1 x_1^j + w_2 x_2^j \right) \right]. \tag{9}$$

The blue expression is linear in w_i, so it can be simplified to $\sum_{j=1}^{n} y^j x_i^j$. For the red expression, we use the chain rule as follows (first we consider a single $j \in [1, n]$)

$$
\begin{aligned}
& \frac{\partial}{\partial w_i} \ln\left[1+\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)\right] && (10)\\
= & \frac{1}{1+\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)} \cdot \frac{\partial}{\partial w_i}\left(1+\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)\right) && (11)\\
= & \frac{1}{1+\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)} \cdot \exp\left(w_0+w_1x_1^j+w_2x_2^j\right) \frac{\partial}{\partial w_i}\left(w_0+w_1x_1^j+w_2x_2^j\right) && (12)\\
= & x_i^j \cdot \frac{\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)}{1+\exp\left(w_0+w_1x_1^j+w_2x_2^j\right)} && (13)
\end{aligned}
$$

2. **[10 Points]** Now use Equation 13 (and the previous discussion) to show that overall, Expression 9, i.e., $\frac{\partial \ell(w)}{\partial w_i}$, is equal to

$$\frac{\partial \ell(w)}{\partial w_i} = \sum_{j=1}^{n} x_i^j (y^j - p(y^j = 1 \mid x^j; w)) \tag{14}$$

Hint: does Expression 13 look like a familiar probability?

Since the log likelihood is concave, it is easy to optimize using gradient ascent. The final algorithm is as follows. We pick a step size η, and then perform the following iterations until the change is $< \epsilon$:

$$w_0^{(t+1)} = w_0^{(t)} + \eta \sum_j \left[y^j - p(y^j = 1 \mid x^j; w^{(t)})\right] \tag{15}$$

$$w_i^{(t+1)} = w_i^{(t)} + \eta \sum_j x_i^j \left[y^j - p(y^j = 1 \mid x^j; w^{(t)})\right]. \tag{16}$$

1.2: General questions about logistic regression

1. **[7 Points]** Explain why logistic regression is a discriminative classifier (as opposed to a generative classifier such as Naive Bayes).

2. **[7 Points]** Recall the prediction rule for logistic regression is if $p(y^j = 1 \mid x^j) > p(y^j = 0 \mid x^j)$, then predict 1, otherwise predict 0. What does the decision boundary of logistic regression look like? Justify your answer (e.g., try to write out the decision boundary as a function of w_0, w_1, w_2 and x_1^j, x_2^j).

Problem 2: Support Vector Machines

Assume we are given a dataset $S = \{((x_1, y_1), \ldots, (x_n, y_n)\}$ of labeled examples with label set $\{1, -1\}$. In this problem, you will derive the SVM algorithm from the large margin principle.

2.1: Hard Margin

Given a linear classifier $f(x) = w^\top x$, which predicts 1 if $f(x) > 0$ and -1 if $f(x) < 0$, recall the definition of the margin,

$$\gamma = \frac{y \cdot f(x)}{||w||_2}.$$

The margin is the distance from a datapoint x to the decision boundary. For the next two problems, assume the data is linearly separable.

1. **[8 Points]** We would like to make this margin γ as large as possible, i.e., maximize the perpendicular distance to the closest point. Thus our objective function becomes

   $$\max_w \min_{i=1}^n \frac{y_i f(x_i)}{||w||_2}.$$

   (Think about why we use this function.) Show that it is equivalent to the following problem:

   $$\min_w \frac{1}{2}||w||_2^2 \quad \text{such that } y_i \cdot (w^\top x_i) \geq 1,\ i = 1, \ldots, n.$$

   (Hint: does it matter if we rescale $w \rightarrow c \cdot w$)?

2. **[7 Points]** If one of the training samples is removed, will the decision boundary shift toward the point removed, shift away from the point removed, or remain the same? Justify your answer.

2.2: Soft Margin

Recall from the lecture notes that if we allow some misclassification in the training data, the SVM optimization is given by

$$\begin{array}{lll} \text{minimize}_{w,\xi_i} & \frac{1}{2}||w||_2^2 + C\sum_{i=1}^n \xi_i & \\ \text{subject to} & y_i(w^\top x_i) \geq 1 - \xi_i & \forall i = 1, \ldots, n \\ & \xi_i \geq 0 & \forall i = 1, \ldots, n \end{array}$$

Recall from the lecture notes $\xi_1, \ldots, \xi_n$ are called slack variables. The optimal slack variables have intuitive geometric interpretation as shown in Figure 1. Basically, when $\xi_i = 0$, the corresponding feature vector $\phi(x_i)$ is correctly classified and it will either lie on the margin of the separator or on the correct side of the margin. Feature vectors with $0 \leq \xi_i \leq 1$ lie within the margin but are still correctly classified. When $\xi_i > 1$, the corresponding feature vector is misclassified. Support vectors correspond to the instances with $\xi_i > 0$ or instances that lie on the margin. The optimal vector w can be represented in terms of $\alpha_1, \ldots, \alpha_n$ as $w = \sum_{i=1}^n \alpha_i y_i \phi(x_i)$.

1. **[8 Points]** Suppose the optimal $\xi_1, \ldots, \xi_n$ have been computed. Use the ξ_i to obtain an upper bound on the number of misclassified instances.

2. **[8 Points]** In the primal optimization of SVM, what is the role of the coefficient C? Briefly explain your answer by considering two extreme cases, $C \rightarrow 0$ and $C \rightarrow \infty$.

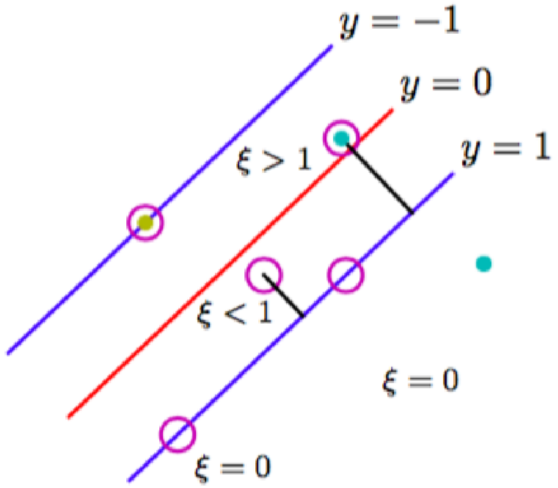

Figure 1: The relationship between the optimal slack variables and the optimal linear separator in the feature space. Support vectors are surrounded with circles.

Problem 3: Kernels

3.1: Kernel computation cost

1. **[7 Points]** Consider we have a two-dimensional input space such that the input vector is $x = (x_1, x_2)^T$. Define the feature mapping $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$. What is the cooresponding kernel function, i.e. $K(x, z)$? Do not leave $\phi(x)$ in your final answer.

2. **[7 Points]** Suppose we want to compute the value of the kernel function $K(x, z)$ from the previous question, on two vectors $x, z \in \mathbb{R}^2$. How many additions and multiplications are needed if you

   (a) map the input vector to the feature space and the perform the dot product on the mapped features?

   (b) compute through the kernel function you derived in question 1?

3.2: Kernel functions

Consider the following kernel function:

$$K(x, x') = \begin{cases} 1, & \text{if } x = x' \\ 0, & \text{otherwise} \end{cases}$$

1. **[7 Points]** Prove this is a legal kernel. That is, describe an implicit mapping $\Phi : X \to \mathcal{R}^m$ such that $K(x, x') = \Phi(x) \cdot \Phi(x')$. (You may assume the instance space X is finite.)

2. **[7 Points]** In this kernel space, any labeling of points in X will be linearly separable. Justify this claim.

3. **[7 Points]** Since all labelings are linearly separable, this kernel seems perfect for learning any target function. Why is this actually a bad idea?

(Bonus) Problem 4: Logistic Regression and M(C)AP

In this question, we assume the same setup as problem 1: we have a training set $\{(x^i, y^i), i = 1, \ldots, n\}$ where $x^i \in \mathbb{R}^2$, and a parametric model

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)} = \frac{\exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)}. \tag{17}$$

1. **(Bonus) [3 points]** Let's look at a MAP estimate for logistic regression.
   Given the Laplace prior $p(w) = \prod_i \frac{1}{2b} e^{\frac{|w_i|}{b}}$, derive the expression to maximize for the M(C)AP estimate. I.e., compute the expression $w^* = \text{argmax}_w \ln[p(w) \prod_j p(y^{(j)}|x^{(j)}, w)]$ (Hint, it should be very similar to Expression 8, but with an extra term corresponding to the prior term.

2. **(Bonus) [3 Points]** What is the expression we should get now for the partial derivative for the M(C)AP estimate? Hint: You should be able to separate out the prior term from the M(C)LE and end up with an expression similar to Expression 14, but with an extra term.

(Bonus) Problem 5: Kernels and Feature Mapping

Consider a binary classification problem in one-dimensional space where the sample contains four data points $S = \{(1, -1), (-1, -1), (2, 1), (-2, 1)\}$ as shown in Figure 2.

1. **[3 Points]** Define $H_t = [t, \infty)$. Consider a class of linear separators $\mathcal{H} = \{H_t : t \in \mathbb{R}\}$, i.e., $\forall H_t \in \mathcal{H}$, $H_t(x) = 1$ if $x \geq t$, otherwise -1. Is there any linear separator $H_t \in \mathcal{H}$ that achieves 0 classification error on this example? If yes, show one of the linear separators that achieves 0 classification error on this example. If not, briefly explain why there cannot be such a linear separator.

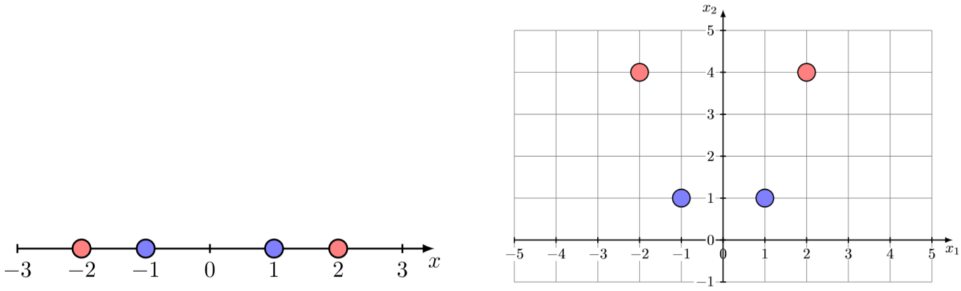

Figure 2: Red points represent instances from class $+1$ and blue points represent instances from class -1. The figure on the left is the original data, and the figure on the right is the data after the feature map transformation.

2. **[3 Points]** Now consider a feature map $\phi : \mathbb{R} \to \mathbb{R}^2$ where $\phi(x) = (x, x^2)$. Apply the feature map to all instances in sample S to generate a transformed sample $S' = \{(\phi(x), y) : (x, y) \in S\}$ shown in Figure 2. Let $\mathcal{H}' = \{ax_1 + bx_2 + c \geq 0 : a^2 + b^2 \neq 0\}$ be a collection of half-spaces in $\mathbb{R}^2$. More specifically, $H_{a,b,c}((x_1, x_2)) = 1$ if $ax_1 + bx_2 + c \geq 0$ otherwise -1. Is there any half-space $H' \in \mathcal{H}'$ that achieves 0 classification error on the transformed sample S'? If yes, give the equation of the max-margin linear separator and compute the corresponding margin. For this question, you can give the equation directly by inspection of Figure 2.

3. **[3 Points]** What is the kernel corresponding to the feature map $\phi(\cdot)$ in the last question, i.e., give the kernel function $K(x, z) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Submission Instructions

You only need to submit the writeup as a pdf file.

Please put these files, together with your writeup, into a folder called hw2, and run the following command

```
$ tar -cvf hw2.tar hw2
```

Then submit your tarfile to Autolab.