Practice questions - Learning

April 25, 2018

1. In class we used a hypothesis of the form $h(x; \mathbf{w}, b) = \sigma(\mathbf{w}^T \phi(x) + b)$, where $\sigma(s) = \frac{1}{1+e^{-s}}$, and the negative log likelihood as a loss function.

$$L(h(x; \mathbf{w}, b), y) = -(y \log h(x; \mathbf{w}, b) + (1 - y) \log(1 - h(x; \mathbf{w}, b))) \quad (1)$$

Consider a different hypothesis class given by $h'(x; \mathbf{w}, b) = \mathbf{w}^T \phi(x) + b$ and a different loss $L'(h'(x; \mathbf{w}, b), y) = \log(1 + e^{-y' h'(x; \mathbf{w}, b)})$, where $y' = 1$ when $y = 1$ and $y' = -1$ when $y = 0$. In this case, h' does not output a probability, it outputs a score that must be high if the true label is 1, and low if the true label is 0. L' is called the *log loss*. Show that $L'(h'(x; \mathbf{w}, b), y) = L(h(x; \mathbf{w}, b), y)$.

Consider two cases.

When $y = 1$:

$$\begin{aligned} L(h(x; \mathbf{w}, b) my) &= -\log h(x; \mathbf{w}, b) && (2)\\ &= -\log \sigma(\mathbf{w}^T \phi(x) + b) && (3)\\ &= -\log \sigma(h'(x; \mathbf{w}, b)) && (4)\\ &= -\log \frac{1}{1 + e^{-h'(x; \mathbf{w}, b)}} && (5)\\ &= \log(1 + e^{-h'(x; \mathbf{w}, b)}) && (6)\\ &= \log(1 + e^{-y' h'(x; \mathbf{w}, b)} && (7) \end{aligned}$$

When $y = 0$: Note that $y' = -1$ in this case.

$$
\begin{aligned}
L(h(x;\mathbf{w},b)my) &= -\log(1 - h(x;\mathbf{w},b)) && (8)\\
&= -\log(1 - \sigma(\mathbf{w}^T\phi(x) + b)) && (9)\\
&= -\log(1 - \sigma(h'(x;\mathbf{w},b))) && (10)\\
&= -\log(1 - \frac{1}{1 + e^{-h'(x;\mathbf{w},b)}}) && (11)\\
&= -\log \frac{e^{-h'(x;\mathbf{w},b)}}{1 + e^{-h'(x;\mathbf{w},b)}} && (12)\\
&= -\log \frac{1}{e^{h'(x;\mathbf{w},b)} + 1} && (13)\\
&= \log(1 + e^{h'(x;\mathbf{w},b)}) && (14)\\
&= \log(1 + e^{-y'h'(x;\mathbf{w},b)} && (15)
\end{aligned}
$$

2. When minimizing a function $F(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ using gradient descent, the update for each iteration is $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \lambda \nabla F(\boldsymbol{\theta})$.

   (a) Is $F(\boldsymbol{\theta}^{(t+1)}) < F(\boldsymbol{\theta}^{(t)}$ always? Do we need a constraint on λ to make sure this is true? Not always: λ should be small enough for the first-order Taylor expansion to hold.

   (b) Will gradient descent always converge to the global minimum of the function? No, it will converge to a local minimum (or more precisely, a critical point: a point with zero gradient)

   (c) What happens if we replace the update rule with $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \lambda \nabla F(\boldsymbol{\theta})$? We will be maximizing the function instead.

3. To compute the bag-of-words descriptor, we need to first run k-means on a dataset of patches to get a set of k-means centers.

   (a) If we are interested in classifying dogs vs cats, which of the following two sets of patches should we use to identify these centers? (a) Patches sampled from car images, (b) Patches sampled from animal images. (b), because they are more related to the task at hand, so we are more likely to identify the right set of k-means centers or "words".

   (b) How does the number of k-means centers affect the dimensionality of the final bag-of-words feature descriptor? The higher the number of centers, the higher the dimensionality.

   (c) How would you change the number of centers to reduce overfitting? Reduce them.