Introduction to Machine Learning, Spring 2025

Homework 3

(Due April 6, 2025 at 11:59pm (CST))

March 18, 2025

1. Please write your solutions in English.
2. Submit your solutions to the course Gradescope.
3. If you want to submit a handwritten version, scan it clearly.
4. Late homeworks submitted within 3 days of the due date will be marked down 25% each day cumulatively Homeworks submitted more than 3 days after the due date will not be accepted unless there is a valid reason, such as a medical or family emergency.
5. You are required to follow ShanghaiTech's academic honesty policies. You are allowed to discuss problems with other students, but you must write up your solutions by yourselves. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious penalties.

1. [15 points] [Perceptron Learning Algorithm] Consider a binary classification problem. The input space is $\mathbb{R}^d$. The output space is $\{+1, -1\}$. For simplicity, we modified the input to be $\mathbf{x} = [x_0, x_1, \cdots, x_d]^\top$ with $x_0 = 1$. The output is predicted using the hypothesis:

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}), \tag{1}$$

where $\mathbf{w} = [w_0, w_1, \cdots, w_d]^\top$ and w_0 is the bias.

The *perceptron learning algorithm* determines $\mathbf{w}$ using a simple iterative method. Here is how it works. At iteration t, where $t = 0, 1, 2, \ldots$, there is a current value of the weight vector, call it $\mathbf{w}(t)$. The algorithm picks an example from $(\mathbf{x}_1, y_1) \cdots (\mathbf{x}_N, y_N)$ that is currently misclassified, call it $(\mathbf{x}(t), y(t))$, and uses it to update $\mathbf{w}(t)$. Since the example is misclassified, we have $y(t) \neq \text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right)$. The update rule is

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t). \tag{2}$$

(a) Show that $y(t)\mathbf{w}^\top(t)\mathbf{x}(t) < 0$. [Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.] [5 points]

(b) Show that $y(t)\mathbf{w}^\top(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^\top(t)\mathbf{x}(t)$. [5 points]

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move "in the right direction". [5 points]

Solution

(a) Since we are considering the misclassified, so we have $y(t) \neq \text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right)$.
And since $y(t), \text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right) \in \{+1, -1\}$, so we have $y(t) \cdot \text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right) = -1 < 0$.
Suppose that $\text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right) = k \cdot \mathbf{w}^\top(t)\mathbf{x}(t)$, where $k > 0$.
So $y(t) \cdot \text{sign}\left(\mathbf{w}^\top(t)\mathbf{x}(t)\right) = y(t) \cdot k \cdot \mathbf{w}^\top(t)\mathbf{x}(t) = k \cdot y(t)\mathbf{w}^\top(t)\mathbf{x}(t) < 0$.
Since $k > 0$, so we have

$$y(t)\mathbf{w}^\top(t)\mathbf{x}(t) < 0$$

So above all, we have proved that $y(t)\mathbf{w}^\top(t)\mathbf{x}(t) < 0$.

(b) Since we are considering the misclassified, so we have $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$.
So

$$y(t)\mathbf{w}^\top(t+1)\mathbf{x}(t) = y(t)\mathbf{w}^\top(t)\mathbf{x}(t) + y(t)y(t)\mathbf{x}^\top(t)\mathbf{x}(t) = y(t)\mathbf{w}^\top(t)\mathbf{x}(t) + y^2(t)\|\mathbf{x}(t)\|^2$$

Since $y(t) \in \{+1, -1\}$, so we have $y^2(t) = 1$.
And since for the simplicity, we have the input to be $\mathbf{x} = [x_0, x_1, \cdots, x_d]^\top$ with $x_0 = 1$, so we have $\mathbf{x} \neq \mathbf{0}$, i.e. $\|\mathbf{x}(t)\|^2 > 0$.
So we have

$$y^2(t)\|\mathbf{x}(t)\|^2 > 0$$

So

$$y(t)\mathbf{w}^\top(t+1)\mathbf{x}(t) = y(t)\mathbf{w}^\top(t)\mathbf{x}(t) + y^2(t)\|\mathbf{x}(t)\|^2 > y(t)\mathbf{w}^\top(t)\mathbf{x}(t)$$

So above all, we have proved that $y(t)\mathbf{w}^\top(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^\top(t)\mathbf{x}(t)$.

(c) We only consider about the misclassified case.
From (a), we knew that

$$y(t)\mathbf{w}^\top(t)\mathbf{x}(t) < 0$$

And from (b), we knew that

$$y(t)\mathbf{w}^\top(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^\top(t)\mathbf{x}(t)$$

So we could see that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is making the $y(t)\mathbf{w}^\top\mathbf{x}(t)$ to the more positive direction, and since if $y(t)\mathbf{w}^\top\mathbf{x}(t) > 0$, then it is a correct classification.
And if the total input data are linearly separable, from what we have learned, we could get that with at most $M = (\frac{R}{\gamma})^2$ such misclassified's movement, where R is the radius of the smallest sphere that contains all the input data, and γ is the margin, then we could get the correct classification.

So above all, we could say that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move "in the right direction".

2. [20 points] [Maximum Margin Classifier] Consider a data set of n d-dimensional sample points, $\{x_1, \ldots, x_n\}$. Each sample point, $x_i \in \mathbb{R}^d$, has a corresponding label, y_i, indicating to which class that point belongs. For now, we will assume that there are only two classes and that every point is either in the given class ($y_i = 1$) or not in the class ($y_i = -1$). Consider the linear decision boundary defined by the hyperplane

$$\mathcal{H} = \left\{x \in \mathbb{R}^d : x \cdot w + \alpha = 0\right\}.$$

The maximum margin classifier maximizes the distance from the linear decision boundary to the closest training point on either side of the boundary, while correctly classifying all training points. Suppose the points are linear seperable, and the margin is γ.

(a) The maximum margin classifier aims to maximize the distance from the training points to the decision boundary. Derive the distance from a point x_i to the hyperplane $\mathcal{H}$. [5 points]

(b) An in-class sample point is correctly classified if it is on the positive side of the decision boundary, and an out-of-class sample is correctly classified if it is on the negative side. Assuming all the points are correctly classified, write a set of n constraints to ensure that all n points are correctly classified. [5 points]

(c) Using the previous parts, write an optimization problem for the maximum margin classifier. For convinent, we should additionally add a constrain $\|w\| = 1$ [5 points]

(d) To simply the optimization problem, we can rewrite the optimization problem in part (c) by setting $w' = \dfrac{w}{\gamma}$ and $\alpha' = \dfrac{\alpha}{\gamma}$. Write the optimization problem for the simlified maximum margin classifier. [5 points]

Solution

(a) For any point x_i, suppose that the projection of x_i on the hyperplane $\mathcal{H}$ is x, then we have

$$x \cdot w + \alpha = 0$$

And since x is the projection of x_i on the hyperplane $\mathcal{H}$, so we have $(x_i - x) \perp \mathcal{H}$, which means $(x_i - x) \parallel w$. So we can suppose that $x_i - x = d\dfrac{w}{\|w\|}$, then we have

$$\begin{aligned}
d\frac{w}{\|w\|} &= x_i - x \\
d\frac{w^\top w}{\|w\|} &= w^\top(x_i - x) \text{ (multiply } w^\top \text{ on both sides)} \\
d\frac{\|w\|^2}{\|w\|} &= w^\top x_i - w^\top x = w^\top x_i + \alpha \; (w^\top x + \alpha = 0) \\
d &= \frac{w^\top x_i + \alpha}{\|w\|}
\end{aligned}$$

And since x_i could be in the positive side or negative side of the hyperplane $\mathcal{H}$, so d may be positive or negetive. So the distance from a point x_i to the hyperplane $\mathcal{H}$ is

$$r = |d| = \frac{|w^\top x_i + \alpha|}{\|w\|}$$

So above all, the distance from a point x_i to the hyperplane $\mathcal{H}$ is

$$r = \frac{|w^\top x_i + \alpha|}{\|w\|}$$

(b) Since all sample points are correctly classified, so for the in-class sample points, the label $y_i = 1$ and it is on the positive side of the decision boundary, so we have $\dfrac{x_i \cdot w + \alpha}{\|w\|} \geq \gamma$. So $\dfrac{y_i(x_i \cdot w + \alpha)}{\|w\|} \geq \gamma$.

For the out-of-class sample points, the label $y_i = -1$ and it is on the negative side of the decision boundary, so we have $\dfrac{-(x_i \cdot w + \alpha)}{\|w\|} \geq \gamma$. So $\dfrac{y_i(x_i \cdot w + \alpha)}{\|w\|} \geq \gamma$.

So above all, we have the constraints as follows:

$$\frac{y_i(x_i \cdot w + \alpha)}{\|w\|} \geq \gamma, \forall i \in \{1, 2, \cdots, n\}$$

It is also acceptible is you write

$$\frac{y_i(x_i \cdot w + \alpha)}{\|w\|} > 0, \forall i \in \{1, 2, \cdots, n\}$$

or

$$y_i(x_i \cdot w + \alpha) > 0, \forall i \in \{1, 2, \cdots, n\}$$

(c) The original problem for the maximum margin classifier is

$$\begin{aligned} \max_{w,\alpha,\gamma} \quad & \gamma \\ \text{subject to} \quad & \|w\| = 1 \\ & y_i(x_i \cdot w + \alpha) \geq \gamma,\ \forall i \in \{1, 2, \cdots, n\} \end{aligned} \tag{3}$$

(d) Let $w' = \dfrac{w}{\gamma}$ and $\alpha' = \dfrac{\alpha}{\gamma}$. And since $\gamma = \dfrac{\|w\|}{\|w'\|} = \dfrac{1}{\|w'\|}$, so maximize γ is equivalent to minimize $\|w'\| = \dfrac{1}{\gamma}$, which has the same effect as minimzing $\|w'\|^2$.
So the original problem is equivalent to

$$\begin{aligned} \min_{w',\alpha'} \quad & \|w'\|^2 \\ \text{subject to} \quad & y_i(x_i \cdot w' + \alpha') \geq 1,\ \forall i \in \{1, 2, \cdots, n\} \end{aligned} \tag{4}$$

3. [10 points] [Leave-one-out Cross-validation]

Select each training example in turn as the single example to be held-out, train the classifier on the basis of all the remaining training examples, test the resulting classifier on the held-out example, and count the errors.

Let the superscript $-i$ denote the parameters we would obtain by finding the SVM classifier f without the ith training example. Define the *leave-one-out CV error* as

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i})), \tag{5}$$

where $\mathcal{L}$ is the zero-one loss. Prove that

$$\text{leave-one-out CV error} \leq \frac{\text{number of support vectors}}{\text{n}} \tag{6}$$

Solution

Then 'Support Vectors' mentioned below are the points that are on the margin of SVM trained by all the training data.
Since we are applying Leave-one-out cross-validation, when leaving the i-th point $\mathbf{x}_i$ to the validation set, we have:

1. If $\mathbf{x}_i$ is not the support vector.
   Then the result of the new SVM would not change compared with the original one, so the error is
   $$\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i})) = 0$$
2. If $\mathbf{x}_i$ is the support vector.
   - If $\mathbf{x}_i$ is the unique support vector.
     Then the result of the SVM would change, and the margin of the new SVM would increase, which means that $\mathbf{x}_i$ would be misclassified, so the error is
     $$\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i})) = 1$$
   - If $\mathbf{x}_i$ is not the unique support vector.
     Then the result of the SVM would not change, so the error is
     $$\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i})) = 0$$

   So combine all the cases, we could get that when $\mathbf{x}_i$ is the support vector, the error $\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i})) \leq 1$.

So we could get that the leave-one-out CV error is

$$\begin{aligned}\text{leave-one-out CV error} &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i}))\\ &\leq \frac{1}{n}\sum_{\mathbf{x}_i\text{is support vector}}\mathcal{L}(y_i, f(\mathbf{x}_i; \mathbf{w}^{-i}, b^{-i}))\\ &\leq \frac{1}{n}\sum_{\mathbf{x}_i\text{is support vector}} 1\\ &= \frac{\text{number of support vectors}}{n}\end{aligned}$$

So above all, we have proved that

$$\text{leave-one-out CV error} \leq \frac{\text{number of support vectors}}{\text{n}}$$

4. [10 points] [Probability and Estimation]

The Poisson distribution is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, the number of packets to arrive in a given time window is often assumed to follow a Poisson distribution. $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}, n > 1$ are i.i.d. samples from exponential distribution with parameter $\lambda > 0$, i.e., $X \sim \text{Expo}(\lambda)$. Recall the PDF of exponential distribution is

$$p(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

(a) To derive the posterior distribution of λ, we assume its prior distribution follows gamma distribution with parameters $\alpha, \beta > 0$, i.e., $\lambda \sim \text{Gamma}(\alpha, \beta)$ (since the range of gamma distribution is also $(0, +\infty)$), thus it's a plausible assumption). The PDF of λ is given by

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$
$$\text{where} \quad \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt, \ \alpha > 0.$$

Show that the posterior distribution $p(\lambda \mid \mathcal{D})$ is also a gamma distribution and identify its parameters. Hints: Feel free to drop constants. [5 points]

(b) Derive the maximum a posterior (MAP) estimation for λ under $\text{Gamma}(\alpha, \beta)$ prior. [5 points]

Solution

(a) From Bayes' Rule, we can get that

$$p(\lambda|\mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$$

Since $\mathcal{D}$ do not contain any λ, so we can get that

$$p(\lambda|\mathcal{D}) \propto p(\mathcal{D}|\lambda)p(\lambda)$$

And since $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$ are i.i.d. samples from exponential distribution with parameter $\lambda > 0$, so we can get that

$$p(\mathcal{D}|\lambda) = p(x_1, x_2, \cdots, x_n|\lambda) = \prod_{i=1}^n p(x_i|\lambda)$$

Since $p(x|\lambda) = \lambda e^{-\lambda x}, x > 0$, so WLOG, we can assume that all the sampling points are positive, i.e. $\forall i, x_i > 0$, then we can get that

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum\limits_{i=1}^n x_i}$$

And since we know that the prior distribution of λ is that $\lambda \sim \text{Gamma}(\alpha, \beta)$, so we can get that

$$p(\lambda) = p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$

So

$$p(\lambda|\mathcal{D}) \propto \lambda^n e^{-\lambda \sum\limits_{i=1}^n x_i} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \propto \lambda^{n+\alpha-1} e^{-\lambda\left(\sum\limits_{i=1}^n x_i + \beta\right)}$$

Since $p(\lambda|\mathcal{D})$ is in terms of conditional probability, so its distribution must be a valid distribution. And from

$$p(\lambda|\mathcal{D}) \propto \lambda^{n+\alpha-1} e^{-\lambda\left(\sum\limits_{i=1}^n x_i + \beta\right)}$$

we can get that the distribution is

$$p(\lambda|\mathcal{D}) \sim \text{Gamma}\left(\alpha + n, \beta + \sum_{i=1}^n x_i\right)$$

So above all, we have proved that the posterior distribution $p(\lambda|\mathcal{D})$ is also a Gamma distribution, and the parameters is that $p(\lambda|\mathcal{D}) \sim \text{Gamma}\left(\alpha + n, \beta + \sum\limits_{i=1}^{n} x_i\right)$

(b) From (a), we get that $p(\lambda|\mathcal{D}) \sim \text{Gamma}\left(\alpha + n, \beta + \sum\limits_{i=1}^{n} x_i\right)$.

and $p(\lambda|\mathcal{D}) \propto \lambda^{n+\alpha-1} e^{-\lambda\left(\sum\limits_{i=1}^{n} x_i + \beta\right)}$.
So the MAP for λ under Gamma(α, β) prior is that

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, \lambda^{\alpha+n-1} e^{-\lambda\left(\beta + \sum\limits_{i=1}^{n} x_i\right)}$$

Take it into the log-likelyhood function, the result of MAP is the same.
So

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}} (\alpha + n - 1) \log \lambda - \left(\beta + \sum_{i=1}^{n} x_i\right) \lambda$$

Let

$$f(\lambda) = (\alpha + n - 1) \log \lambda - \left(\beta + \sum_{i=1}^{n} x_i\right) \lambda$$

then

$$f'(\lambda) = \frac{\alpha + n - 1}{\lambda} - \left(\beta + \sum_{i=1}^{n} x_i\right)$$

And

$$f''(\lambda) = -(\alpha + n - 1) \frac{1}{\lambda^2} < 0$$

So we could find that the function $f(\lambda)$ is a concave function.
So to get the MAP, we need to find the point where the first derivative of $f(\lambda)$ is equal to 0.
i.e.

$$\frac{\alpha + n - 1}{\lambda} - \left(\beta + \sum_{i=1}^{n} x_i\right) = 0$$

So

$$\hat{\lambda}_{MAP} = \frac{\alpha + n - 1}{\beta + \sum\limits_{i=1}^{n} x_i}$$

So above all, the MAP estimation for λ under Gamma(α, β) prior is that

$$\hat{\lambda}_{MAP} = \frac{\alpha + n - 1}{\beta + \sum\limits_{i=1}^{n} x_i}$$

5. [10 points] [Linear Classification] Consider the "Multi-class Logistic Regression" algorithm. Given training set $\mathcal{D} = \{(x^i, y^i) \mid i = 1, \dots, n\}$ where $x^i \in \mathbb{R}^{p+1}$ is the feature vector and $y^i \in \mathbb{R}^k$ is a one-hot binary vector indicating k classes. We want to find the parameter $\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_k] \in \mathbb{R}^{(p+1)\times k}$ that maximize the likelihood for the training set. Introducing the softmax function, we assume our model has the form

$$p(y_c^i = 1 \mid x^i; \beta) = \frac{\exp(\beta_c^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)}$$

where y_c^i is the c-th element of y^i.

(a) Complete the derivation of the conditional log likelihood for our model, which is

$$\ell(\beta) = \ln \prod_{i=1}^{n} p(y_t^i \mid x^i; \beta) = \sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^\top x^i) - y_c^i \ln\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right].$$

For simplicity, we abbreviate $p(y_t^i = 1 \mid x^i; \beta)$ as $p(y_t^i \mid x^i; \beta)$, where t is the true class for x^i. [5 points]

(b) Derive the gradient of $\ell(\beta)$ w.r.t. β_1, i.e.,

$$\nabla_{\beta_1}\ell(\beta) = \nabla_{\beta_1}\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^\top x^i) - y_c^i \ln\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right]$$

Remark: Log likelihood is always concave; thus, we can optimize our model using gradient ascent. (The gradient of $\ell(\beta)$ w.r.t. $\beta_2, \dots, \beta_k$ is similar, you don't need to write them) [5 points]

Solution

(a) Since the model with softmax function is that $p(y_c^i = 1|x^i;\beta) = \dfrac{\exp(\beta_c^\top x^i)}{\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)}$.

And since y^i is a one-hot binary vector, so $y_t^i = 1$, and $\forall c \neq t, y_c^i = 0$.
So $\forall i \in \{1, 2, \cdots, n\}, \forall c \in \{1, 2, \cdots, k\}$, we have

$$p(y^i|x^i;\beta) = p(y_t^i|x^i;\beta) = \prod_{c=1}^{k} p(y_c^i|x^i;\beta)^{y_c^i}$$

So the likelihood is that

$$L(\beta) = \prod_{i=1}^{n} p(y^i|x^i;\beta) = \prod_{i=1}^{n}\prod_{c=1}^{k} p(y_c^i|x^i;\beta)^{y_c^i}$$

So the log-likelihood is that

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n}\sum_{c=1}^{k} y_c^i \log p(y_c^i|x^i;\beta) = \sum_{i=1}^{n}\sum_{c=1}^{k} y_c^i \log \frac{\exp(\beta_c^\top x^i)}{\sum\limits_{c'} exp(\beta_{c'}^\top x^i)}$$

$$= \sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^\top x^i) - y_c^i \log\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right]$$

So above all, the log-likelihood is that

$$\ell(\beta) = \sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^\top x^i) - y_c^i \log\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right]$$

(b) The gradient of $\ell(\beta)$ w.r.t. β_1 is that

$$\nabla_{\beta_1}\ell(\beta) = \nabla_{\beta_1}\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^\top x^i) - y_c^i \log\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right]$$

If $c \neq 1$, then $\nabla_{\beta_1} y_c^i \beta_c^\top x^i = 0$, and if $c = 1$, then $\nabla_{\beta_1} y_c^i \beta_c^\top x^i = y_c^i x^i = y_1^i x^i$.
And $\nabla_{\beta_1} \log\left(\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)\right) = \dfrac{\exp(\beta_1^\top x^i) x^i}{\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)}$.

So

$$\begin{aligned}\nabla_{\beta_1} \ell(\beta) &= \sum_{i=1}^{n} \left(\sum_{c=1}^{k} \nabla_{\beta_1} y_c^i \beta_c^\top x^i - \sum_{c=1}^{k} \nabla_{\beta_1} y_c^i \log\left(\sum_{c'} \exp(\beta_{c'}^\top x^i)\right)\right) \\ &= \sum_{i=1}^{n} \left(y_1^i x^i - \left(\sum_{c=1}^{k} y_c^i\right) \cdot \frac{\exp(\beta_1^\top x^i) x^i}{\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)}\right)\end{aligned}$$

Also, since y^i is a one-hot binary vector, so $y_t^i = 1$, and $\forall c \neq t, y_c^i = 0$.
So $\sum\limits_{c=1}^{k} y_c^i = 1$.
So

$$\nabla_{\beta_1} \ell(\beta) = \sum_{i=1}^{n} \left(y_1^i x^i - \frac{\exp(\beta_1^\top x^i) x^i}{\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)}\right)$$

So above all

$$\nabla_{\beta_1} \ell(\beta) = \sum_{i=1}^{n} \left(y_1^i x^i - \frac{\exp(\beta_1^\top x^i) x^i}{\sum\limits_{c'} \exp(\beta_{c'}^\top x^i)}\right)$$