CS 181 Spring 2018 Section 2 Notes **SOLUTIONS**
(Bayesian Approaches / Linear Classification)

Bayesian Linear Regression

Parameter Distributions

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$. Consider the generative model:

$$y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \beta^{-1}) \tag{1}$$

The likelihood of the data has the form:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}) \tag{2}$$

Put a conjugate prior on the weights (assume precision β^{-1} known):

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{3}$$

We want a posterior distribution on $\mathbf{w}$. Using Bayes' Theorem:

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w}) \tag{4}$$

It turns out that our posterior after n examples is also Gaussian:

$$p(\mathbf{w}|D) = \mathcal{N}(\mathbf{w}|\mathbf{m}_n, \mathbf{S}_n) \tag{5}$$

where

$$\mathbf{S}_n = \left(\mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}\right)^{-1} \tag{6}$$

$$\mathbf{m}_n = \mathbf{S}_n(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{y}) \tag{7}$$

Posterior Predictive Distributions

We have seen how to obtain a posterior distribution over $\mathbf{w}$. But, given this posterior and a new data point $\mathbf{x}^*$, how do we actually make a prediction y^*? How do we deal with *uncertainty* about $\mathbf{w}$? Consider this:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int_{\mathbf{w}} p(y^*|\mathbf{x}^*, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \tag{8}$$

$$= \int_{\mathbf{w}} \mathcal{N}(y^*|\mathbf{w}^\top \mathbf{x}^*, \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_n, \mathbf{S}_n) d\mathbf{w} \tag{9}$$

This is the **posterior predictive** distribution over y^*. This can be interpreted as a weighted average of many predictors, one for each choice of $\mathbf{w}$, weighted by how likely $\mathbf{w}$ is according to the posterior. Since each of the terms on the right hand side follows a normal distribution, we can use some math (see CS181 2017 lecture 5, slide 33) to find that:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \mathcal{N}(y^*|\mathbf{x}^{*\top} \mathbf{m}_n, \mathbf{x}^{*\top} \mathbf{S}_n \mathbf{x}^* + \beta^{-1}) \tag{10}$$

Classification

The goal in classification is to take an input vector $\mathbf{x}$ and assign it to one of K discrete classes C_k, where $k = 1, \ldots, K$. The input space is thus divided into **decision regions** whose boundaries are called **decision boundaries or surfaces**.

Binary Linear Classification

A discriminant function is one that directly assigns each vector $\mathbf{x}$ to a specific class. We first assume two classes, i.e. our responses are binary and $K = 2$. Linear classification seeks to divide the 2 classes by a linear separator in the feature space: if $m = 2$, the separator is a line; if $m = 3$, the separator is a plane; for general m, the separator is a $(m-1)$-dimensional hyperplane. The simplest representation of a linear discriminant function is obtained by taking a linear function of the input vector as such:

$$h(\mathbf{x}; \mathbf{w}, w_0) = \mathbf{w}^\top \mathbf{x} + w_0 \tag{11}$$

The corresponding decision boundary is defined by the relation $h(\mathbf{x}; \mathbf{w}, w_0) = 0$. The classifier will predict $\hat{y} = 1$ if $h(\mathbf{x}; \mathbf{w}, w_0) > 0$, and predict $\hat{y} = -1$ otherwise.

Weight vector $\mathbf{w}$ is orthogonal to every vector lying within the decision surface, and so $\mathbf{w}$ determines the orientation of the decision boundary. Remember the familiar plane equations $c_1 x + c_2 y + c_3 z + c_4 = 0$ for fixed constants $\mathbf{c}$ and variables x, y, z? These are just

dot products with an orthogonality constraint, and thus, they are precisely the kinds of boundaries described here. The input space can also be transformed through a (potentially non-linear) basis function: $h(\mathbf{x}; \mathbf{w}, w_0) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) + w_0,$. This can help with linear separability. More on this soon with neural networks.

Perceptron Algorithm

An important way to train a linear discriminant model is via the perceptron algorithm. This works for a two-class model. Rather than a $0/1$ error function (or sum-squared error), the perceptron algorithm adopts an alternative error function known as hinge loss, which uses the following function :

$$ReLU(z) = \begin{cases} z & z > 0 \\ 0 & o.w. \end{cases} = \max\{0, z\} \tag{12}$$

called a rectified linear activation (ReLU). This function is useful in the following way. Since $h() > 0$ for $y = 1$ and $h() < 0$ for $y = -1$, the product $-h(\mathbf{x}_i; \mathbf{w}, w_0)y_i > 0$ when there is a classification error. The perceptron loss function is defined as:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} ReLU(-h(\mathbf{x}_i; \mathbf{w}, w_0)y_i) \tag{13}$$

$$= - \sum_{i=1:y_i \neq \hat{y}_i}^{n} (\mathbf{w}^\top \mathbf{x}_i + w_0)y_i \tag{14}$$

The first term takes the sum over all training examples of the ReLU function applied to $-h(\mathbf{x}_i; \mathbf{w}, w_0)y_i$. When there is a misclassified example, then this value is positive and it counts as a loss. Equivalently, we can simply write this as the negated sum over all misclassified examples of $h(\mathbf{x}_i; \mathbf{w}, w_0)y_i$.

This loss function has a gradient that is easier to work with than if we had used a $0/1$ error function, and we can now apply stochastic gradient descent. The change in weight vector from step t to $t+1$ is given by the following iteration on an incorrect example:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \frac{\partial}{\partial \mathbf{w}} \mathcal{L}^{(i)}(\mathbf{w}) = \mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i, \tag{15}$$

where η is the learning rate parameter. Note that as the weight vector evolves during training, the set of examples that are misclassified will also change.

Practice Questions

1. **Posterior Weight Distribution By Completing the Square (Bishop 3.7)**

We know from (3.10) in Bishop that the likelihood can be written as

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{w}^\top \mathbf{x}_i, \beta^{-1})$$
$$\propto \exp\left(-\frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})\right)$$

where precision $\beta = \frac{1}{\sigma^2}$ and in the second line above we have ignored the Gaussian normalization constants. By completing the square, show that with a prior distribution on $\mathbf{w}$ given by $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, the posterior distribution $p(\mathbf{w}|D)$ is given by

$$p(\mathbf{w}|D) = \mathcal{N}(\mathbf{w}|\mathbf{m}_n, \mathbf{S}_n)$$

where

$$\mathbf{m}_n = \mathbf{S}_n(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{y})$$
$$\mathbf{S}_n = \left(\mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

Here's the first step. Take ln[(likelihood)(prior)] and collect normalization terms that don't depend on $\mathbf{w}$:

$$\ln p(\mathbf{w}|D) \propto \ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \ln p(\mathbf{w})$$
$$= const - \frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)$$

Hint: Remember, you already know what the posterior should look like. Once you simplify your expression enough, try foiling the posterior in terms of $\mathbf{m}_n$ and $\mathbf{S}_n^{-1}$ and see if you can see the relationship between your expression and this posterior.

Solution:

As the problem statement suggestions, the first step is to take ln[(likelihood)(prior)] and collect normalization terms that don't depend on $\mathbf{w}$:

$$\ln p(\mathbf{w}|D) \propto \ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \ln p(\mathbf{w}) \tag{16}$$

$$= const - \frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \tag{17}$$

Expanding, we have:

$$const - \frac{1}{2}\left(\beta\mathbf{y}^\top\mathbf{y} - \beta\mathbf{y}^\top\mathbf{X}\mathbf{w} - \beta\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \beta\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right) \tag{18}$$

$$+ \left(\mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_0^\top\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{m}_0^\top\mathbf{S}_0^{-1}\mathbf{m}_0\right) \tag{19}$$

Important: Remember that terms like $\mathbf{y}^\top\mathbf{X}\mathbf{w}$ are the same scalar as $\mathbf{w}^\top\mathbf{X}^\top\mathbf{y}$. Collecting together quadratic and linear terms, factoring the $\mathbf{w}$s out, and moving terms that don't depend on $\mathbf{w}$ into the constant, we have

$$const - \frac{1}{2}\left(\mathbf{w}^\top(\mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{w}^\top(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{y})\right) \tag{20}$$

Put aside what we have done so far. Recall that our target looks like:

$$-\frac{1}{2}\left((\mathbf{w} - \mathbf{m}_n)^\top\mathbf{S}_n^{-1}(\mathbf{w} - \mathbf{m}_n)\right), \tag{21}$$

When expanded, this looks like

$$-\frac{1}{2}\left(\mathbf{w}^\top\mathbf{S}_n^{-1}\mathbf{w} - \mathbf{m}_n^\top\mathbf{S}_n^{-1}\mathbf{w} - \mathbf{w}^\top\mathbf{S}_n^{-1}\mathbf{m}_n + \mathbf{m}_n^\top\mathbf{S}_n^{-1}\mathbf{m}_n\right) \tag{22}$$

Drop the term that doesn't have $\mathbf{w}$ and combine the two middle terms

$$-\frac{1}{2}\left(\mathbf{w}^\top\mathbf{S}_n^{-1}\mathbf{w} - 2\mathbf{w}^\top\mathbf{S}_n^{-1}\mathbf{m}_n\right) \tag{23}$$

This looks like what we ended up with in (20) where

$$\mathbf{S}_n^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X}$$

and

$$\mathbf{S}_n^{-1}\mathbf{m}_n = \mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\top\mathbf{y},$$

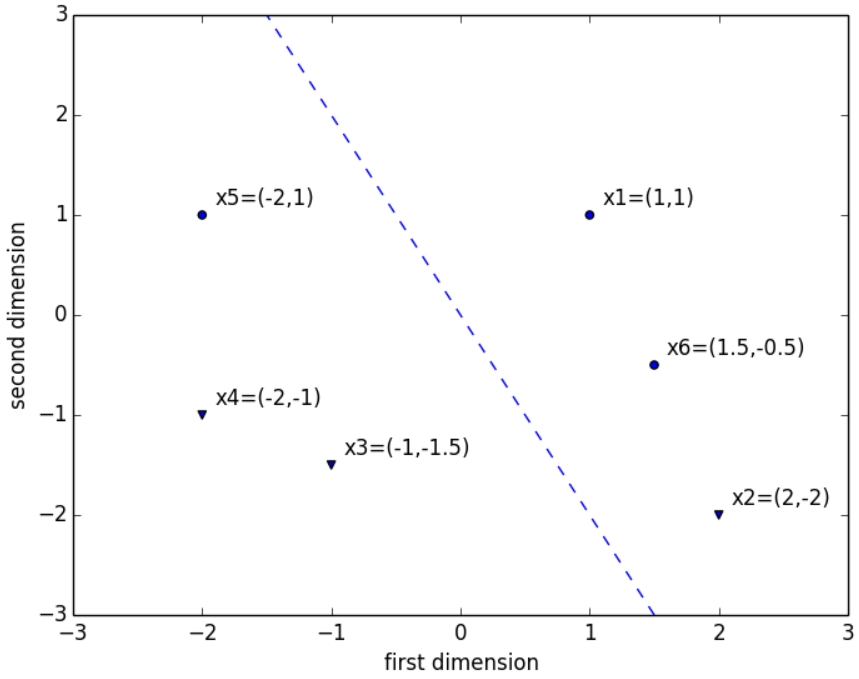
as required. This has the desired form of an (unnormalized) Gaussian.

2. **Small Perceptron Example**

Let's train a perceptron on a small data set. Consider data $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^2$. Let the learning rate $\eta = 0.2$ and let the weights be initialized as:

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, w_0 = 0$$

Let the circles have $y_i = 1$ and the triangles $y_i = -1$. The data and initial separation boundary (determined by $\mathbf{w}$) is illustrated below.



Proceed by iterating over each example until there are no more classification errors. When in doubt, refer to the notes above. We know a priori that we will be able to train the classifier and have no classification errors because one can see visually that the data is linearly separable (note: as mentioned above, if the data were not so obviously linearly separable, a new basis could make it so). How many updates do you have to make? Is this surprising?

Solution:

(a) Consider $\mathbf{x_1}$: $\mathbf{w}^\top \mathbf{x_1} + w_0 = 1 \cdot 1 + 0.5 \cdot 1 + 0 > 0$. This is a correct classification, so we take no action.

(b) Consider $\mathbf{x_2}$: $\mathbf{w}^\top \mathbf{x_2} + w_0 = 1 \cdot 2 + 0.5 \cdot (-2) + 0 > 0$. This is an incorrect classification, so we need to update our weight parameters:

$$\mathbf{w} \leftarrow \mathbf{w} + (0.2)(-1)\begin{pmatrix} 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.9 \end{pmatrix}$$
$$w_0 \leftarrow w_0 + 0.2(-1) = -0.2$$

(c) Consider $\mathbf{x_3}$: $\mathbf{w}^\top \mathbf{x_3} + w_0 = 0.6 \cdot (-1) + 0.9 \cdot (-1.5) + (-0.2) < 0$. This is a correct classification.

(d) Consider $\mathbf{x_4}$: Check that this is a correct classification.

(e) Consider $\mathbf{x_5}$: Check that this is an incorrect classification and our weight parameters are updated to:

$$\mathbf{w} \leftarrow \begin{pmatrix} 0.2 \\ 1.1 \end{pmatrix}$$
$$w_0 \leftarrow 0$$

(f) Consider $\mathbf{x_6}$: Check that this is again an incorrect classification:

$$\mathbf{w} \leftarrow \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$
$$w_0 \leftarrow 0.2$$

All data points are correctly classified now. If the data is linearly separable, we are guaranteed to converge to a solution using the perceptron algorithm in a finite number of steps.

3. **Properties of Softmax**

Consider a K-class classification problem. Let $\{\mathbf{w}_k\}_{k=1}^K$ be defined such that for some data point $\mathbf{x}$, $z_k = \mathbf{w}_k^\top \mathbf{x}$ can be interpretted as a score for $\mathbf{x}$ belonging to class k. Multi-class Logistic Regression (LR) with a trained set of weights assigns $\mathbf{x}$ the class k for which it has the highest such score. The **softmax transformation** takes as input a vector, and outputs a transformed vector of the same size.

$$\text{softmax}(\mathbf{z})_k = \frac{\exp(z_k)}{\sum_{\ell=1}^K \exp(z_\ell)}, \text{ for all } k$$

LR uses the softmax over a vector of K scores $\mathbf{z} = [\mathbf{w}_1^\top \mathbf{x}, \ldots, \mathbf{w}_K^\top \mathbf{x}]$ so that it can be normalized and intepretted as a vector of *probabilities*.

$$p(\mathbf{y} = C_k \mid \mathbf{x}; \{\mathbf{w}_\ell\}_{\ell=1}^K) = \text{softmax}([\mathbf{w}_1^\top \mathbf{x} \ldots \mathbf{w}_K^\top \mathbf{x}]^\top)_k = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{\sum_{\ell=1}^K \exp(\mathbf{w}_\ell^\top \mathbf{x})}.$$

where C_k is a *one-hot* vector with a 1 in coordinate k and 0s elsewhere.
Assuming data $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, the negative log-likelihood can be written as:

$$\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^N \ln p(\mathbf{y}_i \mid \mathbf{x}_i; \{\mathbf{w}_\ell\})$$

The softmax is an important function and you will see it again in other models, such as neural networks. In this problem, we aim to gain intuitions into the properties of softmax and multiclass logistic regression. In the section note solutions, we provide arguments for facts (a), (b), (c), and (d) (make sure you verify these on your own time). Using (d), show that (e) holds:

(a) The output of the softmax is a vector with non-negative components that are at most 1.
(b) The output of the softmax defines a distribution, so the components sum to 1.
(c) Softmax preserves order. This means that if elements $z_k < z_\ell$ in $\mathbf{z}$, then $\text{softmax}(\mathbf{z})_k < \text{softmax}(\mathbf{z})_\ell$ for any k, ℓ.
(d)
$$\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} = \text{softmax}(\mathbf{z})_k (I_{kj} - \text{softmax}(\mathbf{z})_j) \text{ for any } k, j$$
where indicator $I_{kj} = 1$ if $k = j$ and $I_{kj} = 0$ otherwise.
(e) Using (d), show that:
$$\frac{\partial}{\partial \mathbf{w}_j} \mathcal{L}(\{\mathbf{w}_\ell\}) = \sum_{i=1}^N \big[p(\mathbf{y}_i = C_k | \mathbf{x}_i; \{\mathbf{w}_\ell\}) - y_{ik}\big] \mathbf{x}_i$$

Solution:

(a) The j^{th} component of the softmax function $\text{softmax}(\mathbf{z})$ is:

$$\text{softmax}(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_i \exp(z_i)}.$$

As $\exp(x) > 0$ for all $x \in \mathbb{R}$, we have $\exp(z_j) > 0$ and $\sum_i \exp(z_i) > 0$. Thus the output of the softmax function is a vector with non-negative components. Since $\exp(z_j)$ appears in both the numerator and the denominator (as the $i = j$ term in the sum), the denominator must be at least as large as the numerator, and so the components are at most 1.

(b) Summing over the components:

$$\sum_j \text{softmax}(\mathbf{z})_j = \sum_j \frac{\exp(z_j)}{\sum_i \exp(z_i)} = \frac{\sum_j \exp(z_j)}{\sum_i \exp(z_i)} = 1.$$

(c) If $z_j \geq z_k$, then $\exp(z_j) \geq \exp(z_k)$ as the exponential is a monotonically increasing function. Dividing by the positive constant $\sum_i \exp(z_i)$, this inequality implies that:

$$\text{softmax}(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_i \exp(z_i)} \geq \frac{\exp(z_k)}{\sum_i \exp(z_i)} = \text{softmax}(\mathbf{z})_k,$$

which shows that the softmax function preserves the order of the elements of $\mathbf{z}$.

(d) To relate our notation to Bishop (4.106), note that $y_k = \text{softmax}(\mathbf{z})_k$ and $a_j = z_j$.

If $j \neq k$, then:

$$\begin{aligned}\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} &= \frac{\partial}{\partial z_j}\frac{\exp(z_k)}{\sum_i \exp(z_i)} = -\frac{\exp(z_k)}{\left(\sum_i \exp(z_i)\right)^2}\exp(z_j) \\ &= -\frac{\exp(z_k)}{\sum_i \exp(z_i)}\frac{\exp(z_j)}{\sum_i \exp(z_i)} = -\text{softmax}(\mathbf{z})_k \text{softmax}(\mathbf{z})_j.\end{aligned}$$

If $j = k$ then:

$$\begin{aligned}\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} &= \frac{\partial}{\partial z_j}\frac{\exp(z_k)}{\sum_i \exp(z_i)} = \frac{\exp(z_k)}{\sum_i \exp(z_i)} - \frac{\exp(z_j)^2}{\left(\sum_i \exp(z_i)\right)^2} \\ &= \left(1 - \frac{\exp(z_k)}{\sum_i \exp(z_i)}\right)\frac{\exp(z_k)}{\sum_i \exp(z_i)} = \text{softmax}(\mathbf{z})_k(1 - \text{softmax}(\mathbf{z})_j).\end{aligned}$$

Putting these results together: $\boxed{\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} = \text{softmax}(\mathbf{z})_k(I_{kj} - \text{softmax}(\mathbf{z})_j)}$

(e) Write the negative log-likelihood.

$$\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \ln p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})$$

$$\frac{\partial}{\partial \mathbf{w}_j}\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \frac{\partial}{\partial \mathbf{w}_j} \ln p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})$$

Using Derivative of log + chain rule

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})}\right) \frac{\partial}{\partial \mathbf{w}_j} p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})$$

Rewrite using chain rule

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})}\right) \frac{\partial}{\partial z_j} p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\}) \frac{\partial}{\partial \mathbf{w}_j} z_j$$

The derivative at the end is just the derivative of a dot product:

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})}\right) \frac{\partial}{\partial z_j} p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})\mathbf{x}$$

Use the derivative of the softmax found in (d)

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})}\right) \Big(p(\mathbf{y} = C_k|\mathbf{x}_i; \{\mathbf{w}_\ell\})\Big) \Big(I_{kj} - p(\mathbf{y} = C_j|\mathbf{x}_i; \{\mathbf{w}_\ell\})\Big)\mathbf{x}_i$$

Notice that two terms are conveniently reciprocals and simplify

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \Big(I_{kj} - p(\mathbf{y} = C_j|\mathbf{x}_i; \{\mathbf{w}_\ell\})\Big)\mathbf{x}_i$$

Foil the terms

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} I_{kj}\mathbf{x}_i + \sum_{i=1}^{N} p(\mathbf{y} = C_j|\mathbf{x}_i; \{\mathbf{w}_\ell\})\mathbf{x}_i \Big(\sum_{k=1}^{C} y_{ik}\Big)$$

The I_{kj} in the first sum collapses the sum over k to the term where $j = k$. As the $\mathbf{y}_i$ are *one-hot*, we have that $\sum_{k=1}^{K} y_{ik} = 1$. Using these facts:

$$\frac{\partial}{\partial \mathbf{w}_j}\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N} y_{ij}\mathbf{x}_i + \sum_{i=1}^{N} p(\mathbf{y} = C_j|\mathbf{x}_i; \{\mathbf{w}_\ell\})\mathbf{x}_i$$

$$\boxed{= \sum_{i=1}^{N} (p(\mathbf{y} = C_j|\mathbf{x}_i; \{\mathbf{w}_\ell\}) - y_{ij})\mathbf{x}_i}$$