CS 224 - Quiz 2 Review

Fall 2024

Goal

- Recap key topics through a walk-through of 3 practice problems:
  - logistic regression
  - linear regression
  - clustering

Logistic Regression - Problem

[Exercise 10.1 (b) in Murphy's Book*] Let $\mu_{ik} = \text{softmax}(\boldsymbol{\eta}_i)_k$, where $\boldsymbol{\eta}_i = \mathbf{w}^T \mathbf{x}_i$. Show that the gradient of the NLL is given by

$$\nabla_{\mathbf{w}_j} \text{NLL} = \sum_i (\mu_{ij} - y_{ij}) \mathbf{x}_i.$$

*Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

Logistic Regression - Solution

Notations:

- $\mu_{nc} = \text{softmax}(f(\mathbf{x}_n; \mathbf{w}))_c = \text{softmax}(\boldsymbol{\eta}_n)_c = \text{softmax}(\mathbf{w}^\top \mathbf{x}_n)_c$
- $\mathbf{y}_n$ is the one-hot encoding of the label, i.e., $y_{nc} = \mathbb{I}(\mathbf{y}_n = c)$.

Consider a single data point $(\mathbf{x}_i, \mathbf{y}_i)$; let ℓ_i be the log-likelihood of the i-th data, we have

$$\ell_i = \log \left[\prod_{k=1}^{K} (\mu_{ik})^{y_{ik}} \right] = \sum_k y_{ik} \log \mu_{ik}$$

$$\frac{\partial \ell_i}{\partial \mu_{ik}} = \frac{y_{ik}}{\mu_{ik}}, \qquad \frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \mu_{ik}(\delta_{kj} - \mu_{ij}), \qquad \frac{\partial \eta_{ij}}{\partial \mathbf{w}_j} = \frac{\partial \mathbf{w}_j^\top \mathbf{x}_i}{\partial \mathbf{w}_j} = \mathbf{x}_i$$

Logistic Regression - Solution

$$\nabla_{\mathbf{w}_j} \ell_i = \sum_k \frac{\partial \ell_i}{\partial \mu_{ik}} \frac{\partial \mu_{ik}}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \mathbf{w}_j} \qquad \text{(by the chain rule)}$$

$$= \sum_k \frac{y_{ik}}{\mu_{ik}} \mu_{ik} (\delta_{kj} - \mu_{ij}) \mathbf{x}_i = \sum_k y_{ik} (\delta_{kj} - \mu_{ij}) \mathbf{x}_i$$

$$= \sum_k [\mathbb{I}(k = j) y_{ik} \mathbf{x}_i - y_{ik} \mu_{ij} \mathbf{x}_i] = y_{ij} \mathbf{x}_i - \sum_k (y_{ik} \mu_{ij} \mathbf{x}_i)$$

$$= y_{ij} \mathbf{x}_i - \left(\sum_k y_{ik} \right) \mu_{ij} \mathbf{x}_i \qquad (\because \mu_{ij} \text{ and } \mathbf{x}_i \text{ have nothing to do with subscript } k)$$

$$= y_{ij} \mathbf{x}_i - \mu_{ij} \mathbf{x}_i \qquad (\sum_k y_{ik} = 1 \text{ because } \mathbf{y}_i \text{ is one-hot encoded})$$

$$= (y_{ij} - \mu_{ij}) \mathbf{x}_i$$

Logistic Regression - Solution

$$\nabla_{\mathbf{w}_j} \ell_i = (y_{ij} - \mu_{ij})\mathbf{x}_i$$

The negative log-likelihood of the entire dataset: $\mathrm{NLL} = -\sum_i \ell_i$.
Therefore,

$$\nabla_{\mathbf{w}_j} \mathrm{NLL} = -\sum_i (y_{ij} - \mu_{ij})\mathbf{x}_i = \sum_i (\mu_{ij} - y_{ij})\mathbf{x}_i.$$

Linear Regression - Problem

[Exercise 11.2 in Murphy's Book[†]] Assume that $\bar{x} = 0$, so the input data has been centered. Show that the optimizer of

$$J(\mathbf{w}, w_0) = (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbb{I})^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbb{I}) + \lambda\mathbf{w}^T\mathbf{w}$$

is

$$\begin{aligned} \hat{w}_0 &= \bar{y} \\ \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

[†] Murphy, *Probabilistic Machine Learning: An introduction*.

Linear Regression - Solution

$$
\begin{aligned}
J(\mathbf{w}, w_0) &= (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbb{I})^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbb{I}) + \lambda\mathbf{w}^T\mathbf{w} \\
&= \mathbf{y}^T\mathbf{y} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - 2\mathbf{y}^T(\mathbf{x}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w} + (-2w_0\mathbb{I}^T\mathbf{y} + 2w_0\mathbb{I}^T\mathbf{X}\mathbf{w} + w_0\mathbb{I}^T\mathbb{I}w_0)
\end{aligned}
$$

Consider the terms in brackets:

$$
\begin{aligned}
w_0\mathbb{I}^T\mathbf{y} &= w_0 n\bar{y} \\
w_0\mathbb{I}^T\mathbf{X}\mathbf{w} &= w_0\sum_i \mathbf{x}_i^T\mathbf{w} = n\bar{x}^T\mathbf{w} = 0 \\
w_0\mathbb{I}^T\mathbb{I}w_0 &= nw_0^2
\end{aligned}
$$

$$
\implies \quad J(\mathbf{w}, w_0) = \mathbf{y}^T\mathbf{y} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - 2\mathbf{y}^T(\mathbf{x}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w} + (-2w_0 n\bar{y} + nw_0^2)
$$

Linear Regression - Solution

$$J(\mathbf{w}, w_0) = \mathbf{y}^T\mathbf{y} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - 2\mathbf{y}^T(\mathbf{x}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w} + (-2w_0 n\bar{y} + nw_0^2)$$

Optimizing w.r.t. w_0 we find

$$\frac{\partial}{\partial w_0} J(\mathbf{w}, w_0) = -2n\bar{y} + 2nw_0\Big|_{w_0=\hat{w}_0}^{=0}$$
$$\implies \hat{w}_0 = \bar{y}$$

Optimizing w.r.t. $\mathbf{w}$ we find

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, w_0) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w}\Big|_{\mathbf{w}=\hat{\mathbf{w}}}^{=0}$$
$$\implies \hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Clustering - Problem

[From Radford Neal] Suppose that we are fitting a Gaussian mixture model for data items consisting of a single real value, x, using $K = 2$ components. We have $N = 5$ training cases, in which the values of x are 5, 15, 25, 30, 40.
We use the EM algorithm to find the maximum likelihood estimates for the model parameters, which are the mixing proportions for the two components, π_1 and π_2, and the means for the two components, μ_1 and μ_2. The standard deviations for the two components are fixed at 10. Suppose that at some point in the EM algorithm, the E step found that the responsibilities of the two components for the five data items were

$$r_{i1} = [0.2, 0.2, 0.8, 0.9, 0.9], \;\; r_{i2} = [0.8, 0.8, 0.2, 0.1, 0.1]$$

What values for the parameters π_1, π_2, μ_1, and μ_2 will be found in the next M step of the algorithm?

Clustering - Solution

Loss function $J = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)\right)$ is not easily optimized directly w.r.t. all parameters at once, due to the complexity of the sum inside the log.

- **E-Step**: Maximize J w.r.t. responsibilities r_{ik}, where $k \in \{1, 2, \ldots, K\}$.
  - Fix the parameters of each component θ_k (i.e., π_k, μ_k here).
  - Calculate "responsibilities" r_{ik} — how likely the i-th data belongs to the k-th component.
- **M-Step**: Maximize J w.r.t. parameters of each components.
  - Fix the responsibilities r_{ik}.
  - Estimate the optimal value of θ_k.

Clustering - Solution

From the lecture, we know:

$$\pi_k = \frac{1}{N} \sum_i r_{ik}$$

$$\mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

Clustering - Solution

$$\pi_1 = \frac{1}{N}\sum_i r_{i1} = \frac{0.2 + 0.2 + 0.8 + 0.9 + 0.9}{5} = \frac{3}{5} = 0.6$$

$$\pi_2 = \frac{1}{N}\sum_i r_{i2} = \frac{0.8 + 0.8 + 0.2 + 0.1 + 0.1}{5} = \frac{2}{5} = 0.4$$

$$\mu_1 = \frac{\sum_i r_{i1}x_i}{\sum_i r_{i1}} = \frac{0.2 \cdot 5 + 0.2 \cdot 15 + 0.8 \cdot 25 + 0.9 \cdot 30 + 0.9 \cdot 40}{0.2 + 0.2 + 0.8 + 0.9 + 0.9} = \frac{87}{3} = 29$$

$$\mu_1 = \frac{\sum_i r_{i2}x_i}{\sum_i r_{i2}} = \frac{0.8 \cdot 5 + 0.8 \cdot 15 + 0.2 \cdot 25 + 0.1 \cdot 30 + 0.1 \cdot 40}{0.8 + 0.8 + 0.2 + 0.1 + 0.1} = \frac{28}{2} = 14$$