Fundamentals of Machine Learning: Theory

Author: Darien Nouri

Topics Covered

1. Model Selection
2. Gradient of Multi-Class Classifier
3. Maximum Likelihood Estimate of a Gaussian Model
4. Hinge Loss Gradients

Question T1: Model Selection

1. Which i and t should we pick as the best model and why?

   We should pick the model M_t^i with the lowest validation error $\mathcal{L}^i_{val,t}$ with an i based on the lower validation error of either logistic regression ($i = 1$) or SVMs ($i = 0$). We should pick t where the validation error was the lowest for the selected model.

2. How should we report the generalization error of the model?

   We should report the generalization error of the model based on its performance on the test set D_{test} using the parameter configuration that was selected based on validation set performance.

Question T2: Gradient of Multi-Class Classifier

Derivation of the gradient of the cross-entropy loss function $L_w$with respect to the parameter matrix w.

Given,

$$\mathcal{L}_w(x, y) = -\sum_{j=1}^{K} y_j \cdot \log p_j$$

$$p_j = \sigma(w^T x)_j = \frac{e^{w_j^T x}}{\sum_{i=1}^{K} e^{w_i^T x}}$$

$\log(p_j)$ w.r.t. p_j:

$$\frac{\partial}{\partial p_j} \log(p_j) = \frac{1}{p_j}$$

To find the gradient of the cross-entropy loss function $\mathcal{L}_w(x, y)$ we will take its derivative w.r.t. w:

$$\frac{\partial \mathcal{L}_w(x, y)}{\partial w} = \frac{\partial}{\partial w} - \sum_{j=1}^{K} y_j \log p_j$$

$$= -\sum_{j=1}^{K} y_j \frac{\partial \log p_j}{\partial w}$$

Using the chain rule we can find the partial derivatives that compose the gradient such that:

$$\frac{\partial \mathcal{L}_w(x,y)}{\partial w} = -\sum_{j=1}^{K} y_j \frac{\partial \log p_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial w_k^T} \cdot \frac{\partial w_k^T}{\partial w}$$

First we will find the partial derivative of $\mathcal{L}_w(x,y)$ w.r.t P_j:

$$\begin{aligned}
\frac{\partial \mathcal{L}_w(x,y)}{p_j} &= \frac{\partial}{\partial p_j} - \sum_{j=1}^{K} y_i \log p_j \\
&= -\sum_{j=1}^{K} y_i \cdot \log p_j \\
&= -\sum_{j=1}^{K} \frac{y_j}{p_j}
\end{aligned}$$

Next we'll differentiate p_j w.r.t $w_k^T x$. We must consider two cases:

Case 1: $j = k$, to find how the softmax probability of class j changes w.r.t its own score.

$$\begin{aligned}
\frac{\partial p_j}{\partial w_k^T x} &= \frac{\partial}{\partial w_k^T x} \frac{e^{w_j^T x}}{\sum_{i=1}^{K} e^{w_i^T x}} \\
&= \frac{e^{w_j^T x} \sum_{i=1}^{K} e^{w_i^T x} - e^{w_j^T x} e^{w_j^T x}}{(\sum_{i=1}^{K} e^{w_i^T x})^2} \\
&= \frac{e^{w_j^T x} \left(\sum_{i=1}^{K} e^{w_i^T x} - e^{w_j^T x} \right)}{(\sum_{i=1}^{K} e^{w_i^T x})^2} \\
&= \frac{e^{w_j^T x}}{\sum_{i=1}^{K} e^{w_i^T x}} \cdot \left(1 - \frac{e^{w_j^T x}}{\sum_{i=1}^{K} e^{w_i^T x}} \right) \\
&= p_j(1 - p_j)
\end{aligned}$$

Case 2: $j \neq k$, to find how the softmax probability of class j changes w.r.t. the score of a different class k.

$$\frac{\partial p_j}{\partial w_k^T x} = \frac{\partial}{\partial w_k^T x} \frac{e^{w_j^T x}}{\sum_{i=1}^K e^{w_i^T x}}$$

$$= \frac{0 - e^{w_j^T x} e^{w_k^T x}}{(\sum_{i=1}^K e^{w_i^T x})^2}$$

$$= -\frac{e^{w_j^T x}}{\sum_{i=1}^K e^{w_i^T x}} \cdot \frac{e^{w_k^T x}}{\sum_{i=1}^K e^{w_i^T x}}$$

$$= -p_j p_k$$

Now, let's differentiate $w_k^T x$ with respect to w:

$$\frac{\partial w_k^T x}{\partial w} = x$$

$$\begin{aligned}
\frac{\partial \mathcal{L}w(x,y)}{\partial w_k^T} &= -y_k \frac{\partial \log p_k}{\partial p_k}(1-p_k) + \sum_{j \neq k} \frac{y_j}{p_j}(-p_j p_k) \\
&= -y_k(1-p_k) + \sum_{j \neq k} y_j p_k \\
&= -y_k(1-p_k) + p_k \sum_{j \neq k} y_j \\
&= -y_k(1-p_k) + p_k(1-p_k) \\
&= p_k - y_k
\end{aligned}$$

Finally, we find the gradient of $\mathcal{L}_w(x, y)$ with respect to W. Luckily for us that only mean multiplying what we have with the partial derivative of the sarimax input w.r.t w which is simply the input vector x.

$$\frac{\partial \mathcal{L}w(x,y)}{\partial w} = -\sum_{j=1}^{k} y_j \frac{\partial \log p_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial w_k^T} \cdot \frac{\partial w_k^T}{\partial w}$$

$$= x(p_k - y_k)$$

Question T3: Maximum Likelihood Estimate of a Gaussian Model

The pdf of a Gaussian Distribution with mean μ and variance σ^2 is given by:

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The likihood of observing the dataset D with mean μ and variance σ^2:

$$\mathcal{L}(D|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1. Expression of log-likelihood $\mathcal{L}_{\mu,\sigma}(D)$ as a function of μ and σ.

   By taking the logrithm on both sides and expanding:

$$\begin{aligned}
\log \mathcal{L}(D|\mu,\sigma) &= \log \left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right] \\
&= \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{(x-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} - \log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{(x-\mu)^2}{2\sigma^2} \\
&= -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

2. Partial derivative of $\mathcal{L}(D)$ with respect to μ, and equating to zero.

$$\begin{aligned}
\frac{\partial \mathcal{L}(D|\mu,\sigma)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[-nlog(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i-\mu)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)
\end{aligned}$$

Equate to zero:

$$\begin{aligned}
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\
n\mu &= \sum_{i=1}^{n} x_i \\
\mu &= \frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}$$

3. Partial derivative of $\mathcal{L}(D)$ with respect to σ, and equating to zero.

$$\begin{aligned}
\frac{\sigma \mathcal{L}(D|\mu,\sigma)}{\sigma \mu} &= \frac{\sigma}{\sigma \mu}\left[-nlog(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
\frac{n}{\sigma} &= \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\sigma &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}
\end{aligned}$$

Question T4: Hinge loss gradients

Since Hinge loss includes a non-differentiable point at $1 - y \cdot f_\theta(x) = 1$, it is not continuously differentiable at all points. However, the linear segments that compose Hinge loss are differentiable within their intervals. If we have the function

$$L_{\text{Hinge}}(x, y, \theta) = \max\left[0, 1 - y \cdot f_\theta(x)\right]$$

We can reconstruct the function into its piece-wise representation

$$L_{\text{Hinge}}(x, y, \theta) = \begin{cases} 0 & \text{if } y \cdot f_\theta(x) \geq 1, \\ 1 - y \cdot f_\theta(x) & \text{if } y \cdot f_\theta(x) < 1. \end{cases}$$

Such that the gradient of the loss w.r.t θ is

$$\nabla_\theta L_{\text{Hinge}}(x, y, \theta) = \begin{cases} 0 & \text{if } y \cdot f_\theta(x) \geq 1, \\ -y \cdot \nabla_\theta f_\theta(x) & \text{if } y \cdot f_\theta(x) < 1. \end{cases}$$

In the first case, where the loss is 0, when the model's prediction is correct and beyond the margin, defined by $1 - y \cdot f_0(x)$, the parameters are not updated because the example is correctly classified.

In the second case, the loss is the function of θ through $f_\theta(x)$, which is differentiable w.r.t to θ such that its gradient exists.

Therefore, even though there exists a non-differentiable point, namely at $1 - y \cdot f_\theta(x) = 0$, the use of gradient-based optimization is not a problem because:

- when an example is correctly classified the loss and gradient are both zero, and no update is made.
- When an example is misclassified the loss is linear and the gradient can be normally calculated.