Exercise solutions

Linear text classification

1. Let $\boldsymbol{x}$ be a bag-of-words vector such that $\sum_{j=1}^{V} x_j = 1$. Verify that the multinomial probability $\mathrm{p}_{\text{mult}}(\boldsymbol{x}; \boldsymbol{\phi})$, as defined in Equation 2.12, is identical to the probability of the same document under a categorical distribution, $\mathrm{p}_{\text{cat}}(\boldsymbol{w}; \boldsymbol{\phi})$.

2. Suppose you have a single feature x, with the following conditional distribution:

$$\mathrm{p}(x \mid y) = \begin{cases} \alpha, & X = 0, Y = 0 \\ 1 - \alpha, & X = 1, Y = 0 \\ 1 - \beta, & X = 0, Y = 1 \\ \beta, & X = 1, Y = 1. \end{cases} \quad \text{[B.23]}$$

Further suppose that the prior is uniform, $\Pr(Y = 0) = \Pr(Y = 1) = \frac{1}{2}$, and that both $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$. Given a Naïve Bayes classifier with accurate parameters, what is the probability of making an error?

Answer:

$$\hat{y}(X = 0) = 0 \quad \text{[B.24]}$$
$$\hat{y}(X = 1) = 1 \quad \text{[B.25]}$$
$$\Pr(\hat{y} = 0 \mid Y = 1) = \Pr(X = 0 \mid Y = 1) = (1 - \beta) \quad \text{[B.26]}$$
$$\Pr(\hat{y} = 1 \mid Y = 0) = \Pr(X = 1 \mid Y = 0) = (1 - \alpha) \quad \text{[B.27]}$$
$$\Pr(\hat{y} \neq y) = \frac{1}{2}(1 - \beta + 1 - \alpha) \quad \text{[B.28]}$$
$$= 1 - \frac{1}{2}(\alpha + \beta) \quad \text{[B.29]}$$

3. Derive the maximum-likelihood estimate for the parameter $\boldsymbol{\mu}$ in Naïve Bayes.

Answer:

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mathrm{p}_{\text{cat}}(y^{(i)}; \boldsymbol{\mu}) \tag{B.30}$$

$$= \sum_{i=1}^{N} \log \mu_{y^{(i)}} \tag{B.31}$$

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mu_{y^{(i)}} - \lambda \left(\sum_{y=1}^{K} \mu_y - 1 \right) \tag{B.32}$$

$$\frac{\partial \ell(\boldsymbol{\mu})}{\partial \mu_y} = \sum_{i=1}^{N} \frac{\delta\left(y^{(i)} = y\right)}{\mu_y} - \lambda \tag{B.33}$$

$$\mu_y \propto \sum_{i=1}^{N} \delta\left(y^{(i)} = y\right) \tag{B.34}$$

4. The classification models in the text have a vector of weights for each possible label. While this is notationally convenient, it is overdetermined: for any linear classifier that can be obtained with $K \times V$ weights, an equivalent classifier can be constructed using $(K-1) \times V$ weights.

   a) Describe how to construct this classifier. Specifically, if given a set of weights $\boldsymbol{\theta}$ and a feature function $\boldsymbol{f}(\boldsymbol{x}, y)$, explain how to construct alternative weights and feature function $\boldsymbol{\theta}'$ and $\boldsymbol{f}'(\boldsymbol{x}, y)$, such that,

$$\forall y, y' \in \mathcal{Y}, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y') = \boldsymbol{\theta}' \cdot \boldsymbol{f}'(\boldsymbol{x}, y) - \boldsymbol{\theta}' \cdot \boldsymbol{f}'(\boldsymbol{x}, y'). \tag{B.35}$$

   b) Explain how your construction justifies the well-known alternative form for binary logistic regression, $\Pr(Y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}' \cdot \boldsymbol{x})} = \sigma(\boldsymbol{\theta}' \cdot \boldsymbol{x})$, where σ is the sigmoid function.

Answer:

a) Let $\theta_{K,j}$ indicate the weight for base feature j in class K. Then $\theta'_{k,j} = \theta_{k,j} - \theta_{K,j}$, and $f'(\boldsymbol{x}, y) = f(\boldsymbol{x}, y)$ for all $y < K$. This means that $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, K) = 0$.

b) In binary classification, $\boldsymbol{\theta}' = \boldsymbol{\theta}_0 - \boldsymbol{\theta}_1$.

$$\Pr(Y = 0 \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, 0)\right)}{\exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, 0)\right) + \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, 1)\right)} \quad \text{[B.36]}$$

$$= \frac{1}{1 + \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, 1) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, 0)\right)} \quad \text{[B.37]}$$

$$= \frac{1}{1 + \exp\left(-\boldsymbol{\theta}' \cdot \boldsymbol{x}\right)}. \quad \text{[B.38]}$$

5. Suppose you have two labeled datasets D_1 and D_2, with the same features and labels.

- Let $\boldsymbol{\theta}^{(1)}$ be the unregularized logistic regression (LR) coefficients from training on dataset D_1.
- Let $\boldsymbol{\theta}^{(2)}$ be the unregularized LR coefficients (same model) from training on dataset D_2.
- Let $\boldsymbol{\theta}^*$ be the unregularized LR coefficients from training on the combined dataset $D_1 \cup D_2$.

Under these conditions, prove that for any feature j,

$$\theta^*_j \geq \min(\theta^{(1)}_j, \theta^{(2)}_j)$$
$$\theta^*_j \leq \max(\theta^{(1)}_j, \theta^{(2)}_j).$$

6. Let $\hat{\boldsymbol{\theta}}$ be the solution to an unregularized logistic regression problem, and let $\boldsymbol{\theta}^*$ be the solution to the same problem, with L_2 regularization. Prove that $||\boldsymbol{\theta}^*||_2^2 \leq ||\hat{\boldsymbol{\theta}}||_2^2$.

Answer:

Proof. Let the unregularized negative log-likelihood be $\mathcal{L}(\boldsymbol{\theta})$. Let the regularized log-likelihood be $L(\boldsymbol{\theta})$. By assumption, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, so $L(\boldsymbol{\theta}^*) \leq L(\hat{\boldsymbol{\theta}})$.

$$L(\boldsymbol{\theta}^*) \leq L(\hat{\boldsymbol{\theta}}) \quad [B.39]$$
$$\mathcal{L}(\boldsymbol{\theta}^*) + \lambda||\boldsymbol{\theta}^*||_2^2 \leq \mathcal{L}(\hat{\boldsymbol{\theta}}) + \lambda||\hat{\boldsymbol{\theta}}||_2^2 \quad [B.40]$$
$$[B.41]$$

By assumption, $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$, so $\mathcal{L}(\hat{\boldsymbol{\theta}}) \leq \mathcal{L}(\boldsymbol{\theta}^*)$, which implies,

$$\mathcal{L}(\boldsymbol{\theta}^*) + \lambda||\boldsymbol{\theta}^*||_2^2 \leq \mathcal{L}(\boldsymbol{\theta}^*) + \lambda||\hat{\boldsymbol{\theta}}||_2^2 \quad [B.42]$$
$$||\boldsymbol{\theta}^*||_2^2 \leq ||\hat{\boldsymbol{\theta}}||_2^2. \quad [B.43]$$

□

7. As noted in the discussion of averaged perceptron in § 2.3.2, the computation of the running sum $\boldsymbol{m} \leftarrow \boldsymbol{m} + \boldsymbol{\theta}$ is unnecessarily expensive, requiring $K \times V$ operations. Give an alternative way to compute the averaged weights $\overline{\boldsymbol{\theta}}$, with complexity that is independent of V and linear in the sum of feature sizes $\sum_{i=1}^{N} |\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)})|$.

8. Consider a dataset that is comprised of two identical instances $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(2)}$ with distinct labels $y^{(1)} \neq y^{(2)}$. Assume all features are binary, $x_j \in \{0, 1\}$ for all j.

   Now suppose that the averaged perceptron always trains on the instance $(\boldsymbol{x}^{i(t)}, y^{i(t)})$, where $i(t) = 2 - (t \mod 2)$, which is 1 when the training iteration t is odd, and 2 when t is even. Further suppose that learning terminates under the following condition:

$$\epsilon \geq \max_j \left| \frac{1}{t} \sum_t \theta_j^{(t)} - \frac{1}{t-1} \sum_t \theta_j^{(t-1)} \right|. \quad [B.44]$$

   In words, the algorithm stops when the largest change in the averaged weights is less than or equal to ϵ. Compute the number of iterations before the averaged perceptron terminates.

Answer:
Let $\tau = \lceil \frac{t}{2} \rceil$. The weights for a feature which is active for $y^{(1)}$ proceed as: $1, 0, 1, 0, 1, \ldots$. The averaged weight for such a feature proceeds as,

$$\frac{1}{2\tau - 1} \sum_{t=1}^{2\tau-1} \theta^{(t)} = \frac{\tau}{2\tau - 1} \quad \text{[B.45]}$$

$$\frac{1}{2\tau} \sum_{t=1}^{2\tau} \theta^{(t)} = \frac{1}{2}. \quad \text{[B.46]}$$

The algorithm terminates at τ^*, where,

$$\frac{\tau^*}{2\tau^* - 1} - \frac{1}{2} = \epsilon \quad \text{[B.47]}$$

$$\frac{2\tau^*}{2\tau^* - 1} = 2\epsilon + 1 \quad \text{[B.48]}$$

$$2\tau^* = 2\tau^* + 4\epsilon\tau^* - 2\epsilon - 1 \quad \text{[B.49]}$$

$$\tau^* = \frac{2\epsilon + 1}{4\epsilon} = \frac{1}{2} + \frac{1}{4\epsilon} \quad \text{[B.50]}$$

$$t^* = 1 + \frac{1}{2\epsilon} \quad \text{[B.51]}$$

9. Prove that the margin loss is convex in $\boldsymbol{\theta}$. Use this definition of the margin loss:

$$L(\boldsymbol{\theta}) = -\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) + \max_y \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) + c(y^*, y), \quad \text{[B.52]}$$

where y^* is the gold label. As a reminder, a function f is convex iff,

$$f(\alpha x_1 + (1 - \alpha) x_2) \leq \alpha f(x_1) + (1 - \alpha) f(x_2), \quad \text{[B.53]}$$

for any x_1, x_2 and $\alpha \in [0, 1]$.

Answer:

Proof.

$$\begin{aligned}
L(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) =& -\alpha\boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) - (1-\alpha)\boldsymbol{\theta}_2 \boldsymbol{f}(\boldsymbol{x}, y^*) \\
& + \max_y \alpha\boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y) + (1-\alpha)\boldsymbol{\theta}_2 \boldsymbol{f}(\boldsymbol{x}, y^*) + c(y^*, y) && \text{[B.54]} \\
=& \max_y \alpha \left(-\boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) + \boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y) + c(y^*, y)\right) \\
& + (1-\alpha)\left(-\boldsymbol{\theta}_2 \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) + \boldsymbol{\theta}_2 \cdot \boldsymbol{f}(\boldsymbol{x}, y) + c(y^*, y)\right) && \text{[B.55]} \\
\leq& \left[\alpha(-\boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) + \max_y \boldsymbol{\theta}_1 \cdot \boldsymbol{f}(\boldsymbol{x}, y) + c(y^*, y))\right] \\
& + \left[(1-\alpha)(-\boldsymbol{\theta}_2 \cdot \boldsymbol{f}(\boldsymbol{x}, y^*) + \max_y \boldsymbol{\theta}_2 \cdot \boldsymbol{f}(\boldsymbol{x}, y) + c(y^*, y))\right] && \text{[B.56]} \\
\leq& \alpha L(\boldsymbol{\theta}_1) + (1-\alpha) L(\boldsymbol{\theta}_2). && \text{[B.57]}
\end{aligned}$$

The inequality holds because $\max_x f(x) + g(x) \leq \max_x f(x) + \max_{x'} g(x')$: maximizing each term separately yields a sum that is at least as large as finding a single y to maximize the sum jointly. □

Remark Let $\hat{y}_1$ be the maximizer of $L(\boldsymbol{\theta}_1)$, and let $\hat{y}_2$ be the maximizer of $L(\boldsymbol{\theta}_2)$. When $\hat{y}_1 = \hat{y}_2 = y^*$, then $L(\boldsymbol{\theta}_1) = L(\boldsymbol{\theta}_2) = L(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) = 0$. When $\hat{y}_1 = \hat{y}_2 \neq y^*$, then both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are on the linearly decreasing part of the loss function. When $\hat{y}_1 \neq \hat{y}_2$, the inequality is strict.

10. If a function f is m-strongly convex, then for some $m > 0$, the following inequality holds for all x and x' on the domain of the function:

$$f(x') \leq f(x) + (\nabla_x f) \cdot (x' - x) + \frac{m}{2}||x' - x||_2^2. \qquad \text{[B.58]}$$

Let $f(x) = L(\boldsymbol{\theta}^{(t)})$, representing the loss of the classifier at iteration t of gradient descent; let $f(x') = L(\boldsymbol{\theta}^{(t+1)})$. Assuming the loss function is m-convex, prove that $L(\boldsymbol{\theta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t)})$ for an appropriate constant learning rate η, which will depend on m. Explain why this implies that gradient descent converges when applied to an m-strongly convex loss function with a unique minimum.