AI Camp — Math Competition

Time limit: 180 minutes Total score: 100 points

---

1. From likelihood to the logistic (cross-entropy) loss

   Consider a sample of feature vectors $X = \{x_1, \dots, x_\ell\}$ and binary labels $Y = \{y_1, \dots, y_\ell\}$, where $y_i \in \{0, 1\}$. We want to train a linear model with score
   $$z_i = \langle w, x_i \rangle,$$
   by minimizing an empirical risk of the form
   $$Q(w; X, Y) = \sum_{i=1}^{\ell} L(y_i, \langle w, x_i \rangle) \;\to\; \min_{w},$$
   where w is the weight vector and $L(y, z)$ is a smooth loss function.

   In logistic regression we model the probability of class 1 as
   $$\tilde{y}_i := p(y_i = 1 \mid x_i, w).$$
   To measure the quality of such a probabilistic classifier, we use the likelihood $P(Y \mid X, w)$. Assume the pairs (x_i, y_i) are independent across i.

   (a) (4 points) If $y_i \in \{0, 1\}$ and the model predicts $\tilde{y}_i \in [0, 1]$, write the probability of observing y_i as a single expression $p(y_i \mid x_i)$.

   You know that for a Bernoulli label:

   - if $y_i = 1$, then $p(y_i \mid x_i) = \tilde{y}_i$,
   - if $y_i = 0$, then $p(y_i \mid x_i) = 1 - \tilde{y}_i$.

   Use the exponent trick:
   $$p(y_i \mid x_i) = \tilde{y}_i^{\square} (1 - \tilde{y}_i)^{\square},$$
   where the boxes are expressions involving y_i.

   (b) (2 points) Write the joint likelihood $P(Y \mid X, w)$ in terms of the per-example probabilities $p(y_i \mid x_i)$.

   Hint: what is the probability of independent events happening simultaneously?

   (c) (3 points) Take the logarithm of your expression and simplify it. Then write the negative log-likelihood (NLL).

   (d) (3 points) Recall the sigmoid function:
   $$\sigma(z) = \frac{1}{1 + e^{-z}}.$$
   In logistic regression we set
   $$z_i = \langle w, x_i \rangle, \qquad \tilde{y}_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}}.$$
   Rewrite the per-sample negative log-likelihood ℓ_i in terms of y_i and z_i.

   (e) (3 points) Show that ℓ_i can be written in the form
   $$\ell_i = \log(1 + e^{-z_i}) + (\text{something involving } (1 - y_i) z_i).$$

Name: ____________________

(f) (1 point) What simple transformation maps $y_i \in \{0,1\}$ to a label $t_i \in \{-1,1\}$?

(g) (2 points) Find a single expression in t_i and z_i that equals $\log(1+e^{-z_i})$ when $t_i = +1$ and equals $\log(1+e^{z_i})$ when $t_i = -1$.

(h) (2 points) Write the final simplified expression for ℓ_i in terms of t_i and z_i.

2. Logistic regression: separability, MAP, and gradients

Now assume labels are given as $t_i \in \{-1,1\}$. If $p(y=1 \mid x, w) = \sigma(w^\top x)$, we are interested in the probability the model assigns to the correct label.

(a) (2 points) Express the probability of the correct label as a single sigmoid expression in terms of t_i:

$$p(y_i \mid x_i, w) = \sigma(\dots).$$

(b) (1 point) Using the identity above, write the (joint) log-likelihood $\log P(Y \mid X, w)$ and the NLL.

(c) (1 point) Definition: the data are linearly separable if there exists a vector u such that

$$t_i\, u^\top x_i > 0 \quad \text{for all } i.$$

If $t_i u^\top x_i > 0$ for all i, what happens to $t_i(\alpha u)^\top x_i$ as $\alpha \to +\infty$? (Does it go to $+\infty$, $-\infty$, or stay bounded?)

(d) (4 points) Assuming the data are linearly separable, find an upper bound for the log-likelihood. Is this bound achieved by any finite w? Conclude whether the maximum-likelihood estimate exists.

(e) (1 point) Switch from MLE to MAP (maximum a posteriori). Using Bayes' rule, write $P(w \mid X, Y)$ in terms of $P(Y \mid X, w)$ and a prior $P(w)$ (up to a proportionality constant).

(f) (2 points) Introduce L2 regularization and consider the regularized objective

$$J(w) = \text{NLL}(w) + \frac{\lambda}{2}\|w\|^2, \qquad \lambda > 0.$$

Describe what happens to $J(\alpha w)$ as $\alpha \to +\infty$.

(g) (2 points) Compare $J(w)$ to the negative log-posterior. What prior $P(w)$ corresponds to the penalty $\frac{\lambda}{2}\|w\|^2$?

(h) (2 points) Now return to the common $y_i \in \{0,1\}$ form with

$$z_i = w^\top x_i, \qquad \tilde{y}_i = \sigma(z_i).$$

The per-sample negative log-likelihood is

$$\ell_i(w) = -\Big(y_i \log \tilde{y}_i + (1-y_i)\log(1-\tilde{y}_i)\Big).$$

Compute $\frac{d\ell_i}{dz_i}$ and simplify it to an expression involving $\tilde{y}_i$ and y_i.

Hint: you may use

$$\frac{d}{dz}\log\sigma(z) = 1-\sigma(z), \qquad \frac{d}{dz}\log(1-\sigma(z)) = -\sigma(z).$$

(i) (1 point) Using $z_i = w^\top x_i$, what is $\frac{\partial z_i}{\partial w}$?

(j) (2 points) Using the chain rule, show that

$$\nabla_w \mathsf{NLL}(w) = \sum_{i=1}^{\ell} \big(\sigma(w^\top x_i) - y_i\big)\, x_i.$$

(k) (1 point) What is $\nabla_w \big(\frac{\lambda}{2}\|w\|^2\big)$?

(l) (1 point) Combine the previous results to show that

$$\nabla_w J(w) = \sum_{i=1}^{\ell} \big(\sigma(w^\top x_i) - y_i\big)\, x_i + \lambda w.$$

3. Softmax regression and non-identifiability

In multiclass classification, logistic regression generalizes as follows: for each class $k \in \{1, \dots, K\}$ we have a weight vector w_k. The predicted probability is

$$P(y = k \mid x, W) = \frac{e^{\langle w_k, x\rangle}}{\sum_{j=1}^{K} e^{\langle w_j, x\rangle}}.$$

The (softmax) negative log-likelihood for a dataset of size N is

$$L_{\mathsf{sm}}(W) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}[y_i = k]\, \log P(y_i = k \mid x_i, W),$$

where $\mathbb{I}[y_i = k] = 1$ if $y_i = k$ and 0 otherwise.

Let $K = 2$. For simplicity, assume the data are linearly inseparable.

(a) (6 points) Let a be any vector of the same dimension as w_1 and w_2. Define

$$w_1' = w_1 + a, \qquad w_2' = w_2 + a.$$

What happens to

$$P(y = 1 \mid x, W) = \frac{e^{\langle w_1, x\rangle}}{e^{\langle w_1, x\rangle} + e^{\langle w_2, x\rangle}}$$

when you replace w_1, w_2 by w_1', w_2'?

(b) (4 points) What does this invariance imply about whether an optimal solution $W^* = (w_1^*, w_2^*)$ can be unique?

(c) (4 points) Let $v := w_1 - w_2$. Rewrite $P(y = 1 \mid x, W)$ in terms of v only.

(d) (6 points) Show that for $K = 2$ the softmax model reduces to a sigmoid:

$$P(y = 1 \mid x, W) = \sigma\big((w_1 - w_2)^\top x\big).$$

4. Decision trees: optimal constant prediction in a leaf

When building a decision tree, suppose a leaf contains N objects $x_1, \dots, x_N$ with labels $y_1, \dots, y_N$. The prediction in this leaf is a constant $\tilde{y}$. Find the value of $\tilde{y}$ that minimizes each loss:

(a) (6 points) Mean Squared Error (regression):

$$Q = \frac{1}{N}\sum_{i=1}^{N} (y_i - \tilde{y})^2.$$

(b) (7 points) Mean Absolute Error (regression):

$$Q = \frac{1}{N}\sum_{i=1}^{N} |y_i - \tilde{y}|.$$

(c) (7 points) LogLoss (binary classification), with $\tilde{y} \in [0,1]$ and $y_i \in \{0,1\}$:

$$Q = -\frac{1}{N}\sum_{i=1}^{N} \Big(y_i \log \tilde{y} + (1 - y_i) \log(1 - \tilde{y}) \Big).$$

5. Softmax: gradients and convexity

Consider multiclass logistic regression with K classes and weight vectors $w_1, \ldots, w_K \in \mathbb{R}^d$. For an object $x_i \in \mathbb{R}^d$ define the class scores

$$s_{ik} = w_k^\top x_i,$$

and the softmax probabilities

$$p_{ik} = P(y_i = k \mid x_i, W) = \frac{e^{s_{ik}}}{\sum_{j=1}^{K} e^{s_{ij}}}.$$

Let $y_i \in \{1, \ldots, K\}$ be the true class and define one-hot targets $y_{ik} = [y_i = k]$.

(a) (3 points) Write the per-sample negative log-likelihood (cross-entropy) $\ell_i(W)$ using the one-hot targets y_{ik} and the probabilities p_{ik}.

(b) (4 points) Let $\mathsf{LSE}(s_i) = \log\left(\sum_{j=1}^{K} e^{s_{ij}}\right)$. Show that

$$\frac{\partial}{\partial s_{ik}} \mathsf{LSE}(s_i) = p_{ik}.$$

(c) (4 points) Using the previous result, compute $\frac{\partial \ell_i}{\partial s_{ik}}$ and simplify your answer to an expression involving only p_{ik} and y_{ik}.

(d) (3 points) Using $s_{ik} = w_k^\top x_i$, compute the gradient $\nabla_{w_k} \ell_i(W)$.

(e) (2 points) Write $\nabla_{w_k} L(W)$ for the full dataset loss

$$L(W) = \sum_{i=1}^{N} \ell_i(W).$$

(f) (4 points) Let $p_i = (p_{i1}, \ldots, p_{iK})^\top$. Show that the Hessian of ℓ_i with respect to the score vector $s_i = (s_{i1}, \ldots, s_{iK})^\top$ is

$$H_i = \nabla_{s_i}^2 \ell_i = \mathsf{diag}(p_i) - p_i p_i^\top,$$

and argue that H_i is positive semidefinite.