Logistic regression with forgetting

mw, nd

May 8, 2012
Outline

1. $C$ dimensional

2. $(C - 1)$-dimensional - help expand all three methods
Defs and properties

\[ \log_t(ab) = \log_t(a) + \log_t(b) + (1 - t) \log_t(a) \log_t(b) \]
Outline

1. $C$ dimensional

2. $(C - 1)$-dimensional - help expand all three methods
Definition of $G_t$ and $\hat{y}_k$

$G_t : \mathbb{R}^N \rightarrow \mathbb{R}$

$$\sum_{c=1}^{C} \exp_t(\hat{a}_c - G_t(\hat{a})) = 1$$

$$0 = \sum_{c=1}^{C} \frac{\partial}{\partial \hat{a}_k} \exp_t(\hat{a}_c - G_t(\hat{a}))$$

$$= \exp_t(\hat{a}_k - G_t(\hat{a}))^t - \sum_{c=1}^{C} \exp_t(\hat{a}_c - G_t(\hat{a}))^t \frac{\partial G_t(\hat{a})}{\partial \hat{a}_k}$$

$$\iff \frac{\partial G_t(\hat{a})}{\partial \hat{a}_k} = \frac{\exp_t(\hat{a}_k - G_t(\hat{a}))^t}{\sum_{c=1}^{C} \exp_t(\hat{a}_c - G_t(\hat{a}))^t} : = \hat{y}_k$$
Dual of $G$

Since
\[
\frac{\hat{y}_c^{1/t}}{\sum_q \hat{y}_q^{1/t}} = \frac{\exp_t(\hat{a}_c - G_t(\hat{a}))}{\sum_q \exp_t(\hat{a}_q - G_t(\hat{a}))} = \exp_t(\hat{a}_c - G_t(\hat{a}))
\]

\[
\sum_{c=1}^C \hat{y}_c \log_t \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t}} = \hat{a} \cdot \hat{y} - G_t(\hat{a}) = G^*_t(\hat{y})
\]

\[
\frac{\partial G^*_t(\hat{y})}{\partial \hat{y}_k} = \log_t \frac{\hat{y}_k^{1/t}}{\sum_c \hat{y}_c^{1/t}} + \sum_c \frac{\hat{y}_c (\sum_q \hat{y}_q^{1/t})^t}{((\hat{y}_c)^{1/t})^t} \frac{\hat{y}_c^{1/t}}{\sum_q \hat{y}_q^{1/t}} \frac{\partial \hat{y}_c^{1/t}}{\partial \hat{y}_k} = 0
\]

- $G_t$ not strictly convex?, gradient not invertible?
Matching Loss

\[ ML(y, \hat{a}) = \sum_{c=1}^{C} y_c \log_t \frac{y_c^{1/t}}{\sum_c y_c^{1/t}} - \sum_{c} \hat{y}_c \log_t \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t}} \]

\[ - \sum_{c} (y_c - \hat{y}_c) \log_t \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t}} \]

\[ = \sum_{c=1}^{C} y_c \log_t \frac{y_c^{1/t}}{\sum_c y_c^{1/t}} - \sum_{c} y_c \log_t \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t}} \]

\[ = \sum_{c=1}^{C} y_c \log_t \frac{y_c^{1/t}}{\sum_c y_c^{1/t}} + G_t(\hat{a}) - \hat{a} \cdot y \]

\[ = G_t(\hat{a}) - G_t(a) - (\hat{a} - a) \cdot y \quad \text{where } a \in g^{-1}(y) \]

\[ \frac{\partial ML(y, \hat{a})}{\partial \hat{a}_k} \overset{(1)}{=} \frac{\partial G_t(\hat{a})}{\partial \hat{a}_k} - y_k \]

Logistic regression with forgetting

May 8, 2012 7 / 23
Mismatched loss 1

Replace \( \log_t \) in first def. of Matching loss \( ML(\hat{y}, \hat{a}) \) with \( \ln \)

\[
L(y, \hat{a}) = \sum_{c=1}^{C} y_c \ln \frac{y_c^{1/t}}{\sum_{c} y_c^{1/t}} - \sum_{c} y_c \ln \frac{\hat{y}_c^{1/t}}{\sum_{c} \hat{y}_c^{1/t}}
\]

\[
= \sum_{c=1}^{C} y_c \ln \frac{y_c^{1/t}}{\sum_{c} y_c^{1/t}} - \sum_{c} y_c \ln (\exp_t (\hat{a}_c - G_t(\hat{a})))
\]

Derivative

\[
\frac{\partial L(y, \hat{a})}{\partial \hat{a}_k} = - \sum_{c} y_c \exp_t (\hat{a}_c - G_t(\hat{a}))^{t-1} \frac{\partial}{\partial \hat{a}_k} (\hat{a}_c - G_t(\hat{a}))
\]

\[
= - y_k \exp_t (\hat{a}_k - G_t(\hat{a}))^{t-1} + \hat{y}_k \sum_{c} y_c \exp_t (\hat{a}_c - G_t(\hat{a}))^{t-1}
\]
Optimality condition 1

∀ classes $k$ and features $m$

\[
\frac{\partial \sum_n L(y_n, \hat{a}_n)}{\partial w_{k,m}} = 0
\]

\[\iff \sum_n y_{n,k}x_{n,m} \exp_t(\hat{a}_{n,k} - G_t(\hat{a}_n))^{t-1}\]

\[\iff \sum_n \hat{y}_{n,k}x_{n,m} \sum_c y_{n,c} \exp_t(\hat{a}_{n,c} - G_t(\hat{a}_n))^{t-1}\]

\[\iff \sum_n y_{n,k}x_{n,m} \frac{\hat{y}_{n,k}^{1-1/t}}{\hat{y}_{n,[y_n]}^{1-1/t}} = \sum_n \hat{y}_{n,k}x_{n,m} \sum_c y_{n,c} \hat{y}_{n,c}^{1-1/t}\]

\[y_n \text{ unit} \iff \sum_{n: [y_n] = k} x_{n,m} \frac{\hat{y}_{n,k}^{1-1/t}}{\hat{y}_{n,[y_n]}^{1-1/t}} = \sum_n \hat{y}_{n,k}x_{n,m} \hat{y}_{n,[y_n]}^{1-1/t}\]
Mismatched loss 2

Start with matching loss for log but replace log with $\log_t$

$$L(y, \hat{a}) = \sum_c (y_c \log_t y_c - y_c \log_t \hat{y}_c)$$

where $\hat{y}_c = \frac{\exp(\hat{a}_c)}{\sum_c \exp(\hat{a}_c)} = \exp(\hat{a}_c - G(\hat{a}))$ and $G(\hat{a}) = \ln \sum_c \exp \hat{a}_c$

$$\frac{\partial L(y, \hat{a})}{\partial \hat{a}_k} = - \sum_c y_c \hat{y}_c^{-t} \frac{\partial}{\partial \hat{a}_k} \underbrace{\exp(\hat{a}_c - G(\hat{a}))}_{\hat{y}_c}$$

$$= - \sum_c y_c \hat{y}_c^{-t} \hat{y}_c \frac{\partial}{\partial \hat{a}_k} (\hat{a}_c - G(\hat{a}))$$

$$= - y_k \hat{y}_k^{1-t} + \hat{y}_k \sum_c y_c \hat{y}_c^{1-t}$$
\[ \forall \text{ classes } k \text{ and features } m \]

\[
\frac{\partial}{\partial w_{k,m}} \sum_n L(y_n, \hat{a}_n) = 0
\]

\[
\iff \sum_n y_{n,k} x_{n,m} \hat{y}_{n,k}^{1-t} = \sum_n \hat{y}_{n,k} x_{n,m} \sum_c y_{n,c} \hat{y}_{n,c}^{1-t}
\]

\[
y_{n \text{ unit}} \iff \sum_{n: [y_n] = k} x_{n,m} \hat{y}_{n,k}^{1-t} = \sum_n \hat{y}_{n,k} x_{n,m} \hat{y}_{n,[y_n]}^{1-t}
\]

Rewrites

\[
\hat{y}_c^{1-t} = \frac{\exp((1 - t)\hat{a}_c)}{(\sum_{c'} \exp(\hat{a}_{c'}))^t - 1} = 1 - (1 - t)(- \log_t \hat{y}_c)
\]

\[
\sum_c y_c \hat{y}_c^{1-t} = 1 - (1 - t)(- \sum_c y_c \log_t \hat{y}_c)
\]
Compare Mismatched Loss 1 versus 2

- How are $\hat{y}^{1-1/t_1}$ and $\hat{y}^{1-t_2}$ related?
- $t_1 \in [1..2]$ and $t_2 \in [0..1]$?
- $t_1 \approx 1/t_2$?
Energy based method 3

\[ L(y, \hat{a}) = \sum_c y_c \log y_c - y_c \ln \left( \frac{\exp(s(\hat{a}_c))}{\sum_c \exp(s(\hat{a}_c))} \right) \]

\[ = \sum_c y_c \log y_c - \sum_c y_c s(\hat{a}_c) + \ln(\sum_c \exp(s(\hat{a}_c))) \]

Derivative

\[ \frac{\partial L(y, \hat{a})}{\partial \hat{a}_k} = (\hat{y}_k - y_k) s'(\hat{a}_k) \]

Optimality condition 3

\[ \frac{\partial \sum_n L(y_n, \hat{a}_n)}{\partial w_{k,m}} = 0 \]

\[ \iff \sum_n \hat{y}_{n,k} x_{n,m} s'(\hat{a}_{n,k}) = \sum_n y_{n,k} x_{n,m} s'(\hat{a}_{n,k}) \]
Robustness

Must contain a tension between two parts

- trust the good
- mistrust the bad
Outline

1. $C$-dimensional

2. $(C - 1)$-dimensional - help expand all three methods
Collapse to $C - 1$ dimensions

For $1 \leq c \leq C - 1$

- $\tilde{w}_c := w_c - w_C$, yielding $\tilde{a}_c = \hat{a}_c - \hat{a}_C$.
- $\tilde{G}_t(\tilde{a}) := G_t(\hat{a}) - \hat{a}_C$
- $y_C = 1 - \sum_{c=1}^{C-1} y_c$ and $\hat{y}_C = 1 - \sum_{c=1}^{C-1} \hat{y}_c$

$$\sum_{c=1}^{C} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a})) = 1 \iff \sum_{c=1}^{C-1} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a})) + \exp_t(-\tilde{G}_t(\tilde{a})) = 1$$

- Define equivalence class: $\hat{a} \sim \hat{a}'$ if $\hat{y}(\hat{a}) = \hat{y}(\hat{a}')$
  Claim: $\hat{a} \sim \hat{a}'$ iff $\hat{a}' = \hat{a} + q \mathbf{1}$, for $q \in \mathbb{R}$ and $\mathbf{1} = (1, \ldots, 1)^\top$
  Each class contains one vector $\hat{a}$ with $\hat{a}_C = 0$

- $\tilde{G}_t^*$ are strictly convex and therefore $\tilde{G}_t$ is so as well
  Gradient of $\tilde{G}_t^*$ is invertible and has closed form
  Gradient of $\tilde{G}_t$ is invertible but no closed form?
Definition of \( \tilde{G}_t \) and \( \tilde{y}_k \)

\[
\tilde{G}_t : \mathbb{R}^{N-1} \rightarrow \mathbb{R}
\]

\[
\left( \sum_{c=1}^{C-1} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a})) \right) + \exp_t(-\tilde{G}_t(\tilde{a})) = 1
\]

\[
0 = \sum_{c=1}^{C-1} \frac{\partial}{\partial \tilde{a}_k} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a})) + \frac{\partial}{\partial \tilde{a}_k} \exp_t(-\tilde{G}_t(\tilde{a}))
\]

\[
= \exp_t(\tilde{a}_k - \tilde{G}_t(\tilde{a}))^t - \sum_{c=1}^{C-1} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a}))^t \frac{\partial \tilde{G}_t(\tilde{a})}{\partial \tilde{a}_k} - \exp_t(\tilde{G}_t(\tilde{a}))^t \frac{\partial \tilde{G}_t(\tilde{a})}{\partial \tilde{a}_k}
\]

\[
\frac{\partial \tilde{G}_t(\tilde{a})}{\partial \tilde{a}_k} = \frac{\exp_t(\tilde{a}_k - \tilde{G}_t(\tilde{a}))^t}{\left( \sum_{c=1}^{C-1} \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a}))^t \right) + \exp_t(\tilde{G}_t(\tilde{a}))^t} := \tilde{y}_k
\]
Dual of $\tilde{G}$

Since

$$\frac{\tilde{y}_{c=1}^{1/t}}{\sum_c^{C-1} \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}} = \exp_t(\tilde{a}_c - \tilde{G}_t(\tilde{a}))$$

we can derive the dual $\tilde{G}_t^*$ of $\tilde{G}_t$

$$- \sum_{c=1}^{C-1} \tilde{y}_c \log_t \frac{\tilde{y}_c^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}}$$

$$- (1 - \sum_c \tilde{y}_c) \log_t \frac{(1 - \sum_c \tilde{y}_c)^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}}$$

$$= \tilde{G}_t(\tilde{a}) - \tilde{a} \cdot \tilde{y} = \tilde{G}_t^*(\tilde{y})$$
Gradient of $\tilde{G}^*$ has closed form and is invertible

As shown in genent.tex by Nan Ding

$$[g^{-1}(\tilde{y})]_k = \frac{\partial \tilde{G}^*_t(\tilde{y})}{\partial \tilde{y}_k}$$

$$= \log_t \frac{\tilde{y}_k^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}} - \log_t \frac{(1 - \sum_c \tilde{y}_c)^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}}$$

By plugging in $\tilde{y}$, we see that $[g^{-1}(\tilde{y})]_k = \tilde{a}_k - \tilde{G}(\tilde{a}) - (-\tilde{G}(\tilde{a})) = \tilde{a}_k$

Since $g^{-1}(\tilde{y})$ invertible, $\tilde{G}^*$ should be convex?

Therefore $\tilde{G}$ convex as well
(C − 1)-dimensional Matching Loss

\[ ML(y, \hat{a}) = \]

\[ \sum_{c=1}^{C-1} y_c \log_t \frac{y_c^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}} + \left(1 - \sum_c y_c\right) \log_t \frac{(1 - \sum_c y_c)^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}} \]

\[ - \sum_{c=1}^{C-1} \tilde{y}_c \log_t \frac{\tilde{y}_c^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}} - \left(1 - \sum_c y_c\right) \log_t \frac{(1 - \sum_c \tilde{y}_c)^{1/t}}{\sum_c \tilde{y}_c^{1/t} + (1 - \sum_c \tilde{y}_c)^{1/t}} \]

\[ = \sum_{c=1}^{C-1} y_c \log_t \frac{y_c^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}} + \left(1 - \sum_c y_c\right) \log_t \frac{(1 - \sum_c y_c)^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}} \]

\[ + \tilde{G}_t(\tilde{a}) - \tilde{y} \cdot \tilde{a} \]

\[ = \tilde{G}_t(\tilde{a}) - \tilde{G}_t(a) - (\tilde{a} - a) \cdot y \quad \text{where} \quad a = g^{-1}(y) \in \mathbb{R}^{N-1} \]

\[ \frac{\partial ML(y, \hat{a})}{\partial \hat{a}_k} = \left\{ \frac{\partial \tilde{G}_t(\tilde{a})}{\partial \hat{a}_k} \right\}_{\tilde{y}_k} - y_k \]
Mismatched loss 1

Replace $\log_t$ in 1st form of Maching Loss with \ln

$$L(y, \tilde{a}) = \sum_{c=1}^{C-1} y_c \ln \frac{y_c^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}} + (1 - \sum_c y_c) \ln \frac{(1 - \sum_c y_c)^{1/t}}{\sum_c y_c^{1/t} + (1 - \sum_c y_c)^{1/t}}$$

$$- \sum_{c=1}^{C-1} y_c \ln \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t} + (1 - \sum_c \hat{y}_c)^{1/t}} - (1 - \sum_c y_c) \ln \frac{(1 - \sum_c \hat{y}_c)^{1/t}}{\sum_c \hat{y}_c^{1/t} + (1 - \sum_c \hat{y}_c)^{1/t}}$$

$$= \sum_{c=1}^{C-1} y_c \ln \frac{\hat{y}_c^{1/t}}{\sum_c \hat{y}_c^{1/t} + (1 - \sum_c \hat{y}_c)^{1/t}} + (1 - \sum_c y_c) \ln \frac{(1 - \sum_c \hat{y}_c)^{1/t}}{\sum_c \hat{y}_c^{1/t} + (1 - \sum_c \hat{y}_c)^{1/t}}$$

$$- \sum_{c=1}^{C-1} y_c \ln (\exp_t (\tilde{a}_c - \tilde{G}(\tilde{a}))) - (1 - \sum_{c=1}^{C-1} y_c) \ln (\exp_t (-\tilde{G}(\tilde{a})))$$
$\frac{\partial L(y, \tilde{a})}{\partial \tilde{a}_k} = - \sum_c y_c \exp_t (\tilde{a}_c - \tilde{G}_t(\tilde{a}))^{t-1} \frac{\partial}{\partial \tilde{a}_k} (\tilde{a}_c - \tilde{G}_t(\tilde{a})) - (1 - \sum_c y_c) \exp_t (-\tilde{G}_t(\tilde{a}))^{t-1}$

$= -y_k \exp_t (\tilde{a}_k - \tilde{G}_t(\tilde{a}))^{t-1}$

$+ \tilde{y}_k \left( \sum_c y_c \exp_t (\tilde{a}_c - \tilde{G}_t(\tilde{a}))^{t-1} + (1 - \sum_c y_c) \exp_t (-\tilde{G}_t(\tilde{a}))^{t-1} \right)$
Mismatched loss 2

\[
L(y, \tilde{a}) = \sum_{c=1}^{C-1} (y_c \log_t y_c + (1 - \sum_c y_c) \log_t (1 - \sum_c y_c)) \\
- y_c \log_t \hat{y}_c - (1 - \sum_c y_c) \log_t (1 - \sum_c \hat{y}_c))
\]

where \( \hat{y}_c = \frac{\exp(\tilde{a}_c)}{1+\sum_c \exp(\tilde{a}_c)} = \exp(\tilde{a}_c - \tilde{G}(\tilde{a})) \) and \( \tilde{G}(\tilde{a}) = \ln(1 + \sum_c \exp \tilde{a}_c) \)
\[
\frac{\partial L(y, \tilde{a})}{\partial \tilde{a}_k} = - \sum_c y_c \hat{y}_c^{1-t} \frac{\partial}{\partial \tilde{a}_k} \left( \hat{y}_c \exp(\tilde{a}_c - \tilde{G}(\tilde{a})) \right) \\
- (1 - \sum_c y_c) (1 - \sum_c \hat{y}_c)^{-t} \frac{\partial}{\partial \tilde{a}_k} \left( 1 - \sum_c \hat{y}_c \right) \exp(-\tilde{G}(\tilde{a})) \\
= - \sum_c y_c \hat{y}_c^{1-t} \hat{y}_c \frac{\partial}{\partial \tilde{a}_k} (\tilde{a}_c - \tilde{G}(\tilde{a})) \\
- (1 - \sum_c y_c) (1 - \sum_c \hat{y}_c)^{-t}(1 - \sum_c \hat{y}_c) \frac{\partial}{\partial \tilde{a}_k} (-\tilde{G}(\tilde{a})) \\
= -y_k \hat{y}_k^{1-t} + \hat{y}_k \left( \sum_c y_c \hat{y}_c^{1-t} + (1 - \sum_c y_c)(1 - \sum_c \hat{y}_c)^{t-1} \right)
\]
Synopsis

- Mismatched loss 3 does not collapse to $C - 1$ dimensions
- Only matching loss seems to collapse
- None of the 3 mismatched losses collapse