Optimization Methods in Machine Learning
Lecture 5-6: Logistic Loss and regularization

Katya Scheinberg

Lehigh University

Spring 2016

The consequence of 01 loss

- Consider the minimization the empirical error problem subject to some constraints.

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, z\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \text{loss}_{01}(z_i, y_i)$$
$$\text{s.t. } z_i = w^T x_i + b$$

- z is the output of the predictor for x_i
- We care about the sign of y and z.

$$\text{loss}_{01} = \begin{cases} 0 & \text{if } y_i z_i > 0 \\ 1 & \text{if } y_i z_i \leq 0 \end{cases}$$

- If $z = 0$, it incurs a loss of 1 since the prediction can never match the true label y.

The consequence of 01 loss

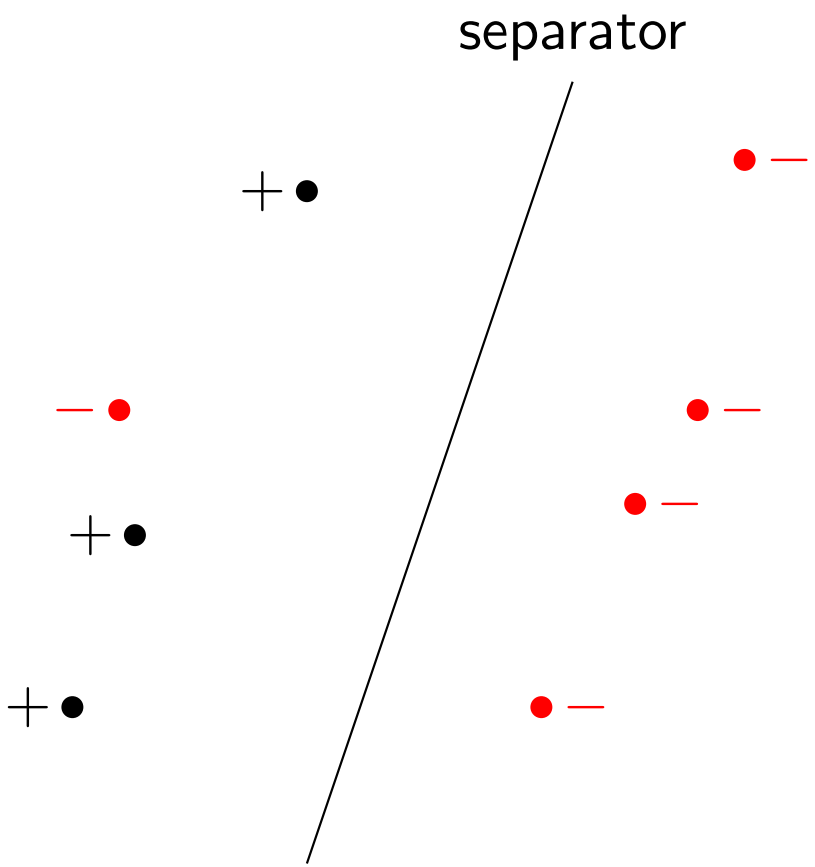
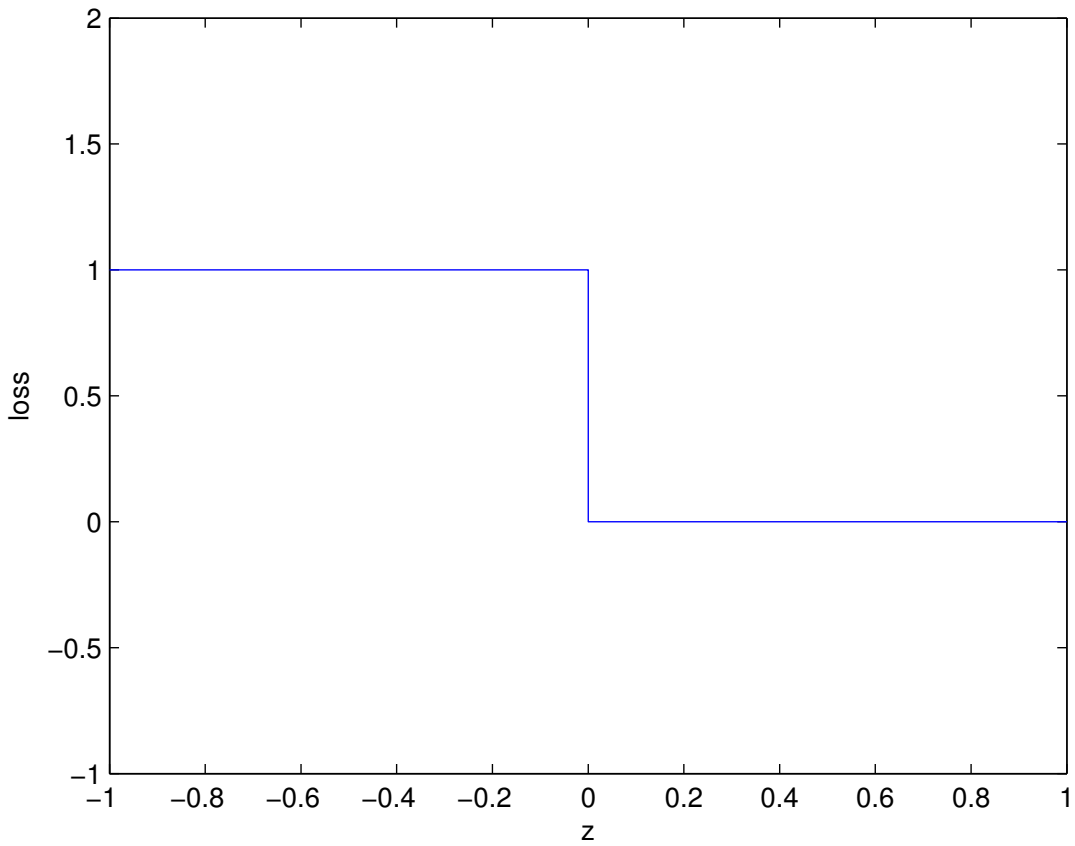


Figure: Affine separator performance in the case of 01 loss. 01 loss is not good at measuring the loss magnitude.

The consequence of 01 loss

- 01 loss function is not convex



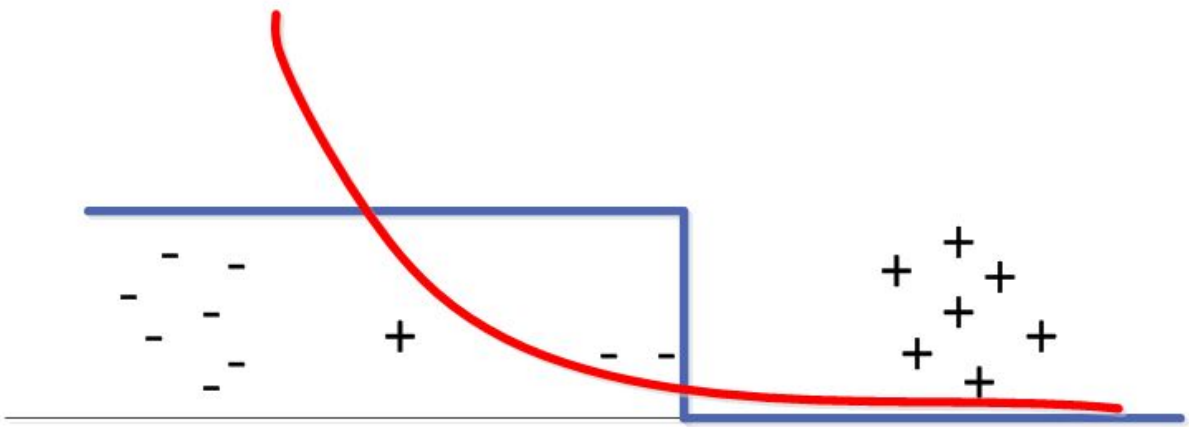
- At 0, the Lipschitz constant is ∞

Logistic Loss

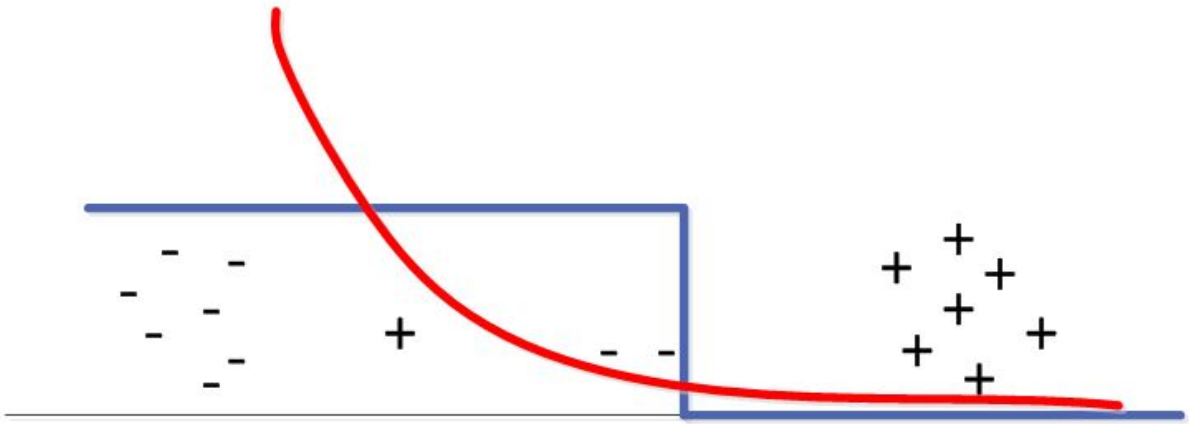
Consider a hypothesis class $\mathcal{H}$ of linear functions $h(x) = w^T x + b$ of size m, and its logistic loss.

$$\min_{w,b} \frac{1}{m} \sum_i \log(1 + e^{-y_i(w^T x_i + b)}).$$

Logistic loss minimizer may give a different separation from 01 loss minimizer.



Logistic Loss



from the example, we can see logistic loss have some advantages over 01 loss:

- Convex function
- Lipschitz function

Question: Is logistic loss good enough?

Behavior of the smooth loss function

$$\min_w \; f(w) = \frac{1}{m} \sum_{i=1}^{m} \text{loss}_g(h(w, x_i), y_i)$$

- If y_i is $+1$, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ is a monotonically decreasing function of $h\left(w, x_i\right)$.
  - That is, for $y_i = +1$, when $h\left(w, x_i\right)$ increases, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ decreases monotonically.
  - A stronger correct prediction leads to lower loss.
- If y_i is -1, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ is a monotonically increasing function of $h\left(w, x_i\right)$.
  - That is, for $y_i = -1$, when $h\left(w, x_i\right)$ increases, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ increases monotonically.
  - A stronger incorrect prediction leads to higher loss.

Behavior of the smooth loss function

$$\min_w \; f(w) = \frac{1}{m} \sum_{i=1}^{m} \text{loss}_g(h(w, x_i), y_i)$$

- If y_i is $+1$, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ is a monotonically decreasing function of $h\left(w, x_i\right)$.
  - That is, for $y_i = +1$, when $h\left(w, x_i\right)$ increases, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ decreases monotonically.
  - $h(w, x)$ should be a **concave** function in w for $f(w)$ to be convex.
- If y_i is -1, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ is a monotonically increasing function of $h\left(w, x_i\right)$.
  - That is, for $y_i = -1$, when $h\left(w, x_i\right)$ increases, $\text{loss}_g\left(h\left(w, x_i\right), y_i\right)$ increases monotonically.
  - $h(w, x)$ should be a **convex** function in w for $f(w)$ to be convex.

Logistic loss and linear predictors

Conclustion: $h(w, x)$ has to be **linear** in w!

$$\min_{w,b} \; f(w) = \frac{1}{m} \sum_{i=1}^{m} \text{loss}_g(h(x_i), y_i) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i + b)})$$

Different mapping $\phi(x_i)$

$$\mathcal{H}: \; x \mapsto w^T \phi(x)$$

or

$$\mathcal{H} = \left\{ h_w(\cdot) \mid h_w(x) = w^T \phi(x), \; x \in \mathbb{R}^D, \; w \in \mathbb{R}^D \right\}.$$

If $\mathcal{X} = \mathbb{R}^2$, $x \in \mathcal{X}$,
$\phi(\cdot) \; : \; \mathcal{X} \to \mathbb{R}^3$, so that now $D = 3$

$$\phi(x) = \begin{bmatrix} x[1] & x[2] & 1 \end{bmatrix}^T$$
$$w = \begin{bmatrix} w_1 & w_2 & b \end{bmatrix}^T$$

If $\mathcal{X} = \mathbb{R}^2$, $x \in \mathcal{X}$
$\phi_2(\cdot) \; : \; \mathcal{X} \to \mathbb{R}^6$, so that now $D = 6$

$$\phi_2(x) = \begin{bmatrix} x[1] & x[2] & (x[1])^2 & (x[2])^2 & x[1]\,x[2] & 1 \end{bmatrix}^T$$
$$w = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & b \end{bmatrix}^T$$

Logistic loss and regularization

Consider the case when $\hat{w}^T x$ separates the data without error. What does this mean?

$$y_i(\hat{w}^T x_i) > 0, \quad \forall i = 1, \ldots m$$

The consider empirical risk minimizer

$$\min_w \; f(w) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i)})$$

Where is the minimum attained?

Logistic loss and regularization

Consider the case when $\hat{w}^T x$ separates the data without error. What does this mean?

$$y_i(\hat{w}^T x_i) > 0, \quad \forall i = 1, \dots m$$

The consider empirical risk minimizer

$$\min_w \ f(w) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i)})$$

Where is the minimum attained? Consider $w = 100\hat{w}$... or $w = 1000000\hat{w}$.
Clearly as $w \to \infty$ $f(w) \to 0$. The means that the problem is not well defined. Also this means that we need to control $\|w\|$.

An Example

For example, $w^T x_i = 0.001$, then the logarithm term could be:

$$\log(1 + e^{-0.001}) \approx \log 2 = 0.301.$$

Without changing the hyperplane, we can multiple scalar on both side of $w^T x_i = 0.001$, say $\bar{w}^T x_i = 1$, with $\bar{w}^T = 1000 w^T$. Then

$$\log(1 + e^{-1}) \approx \log 1.36 = 0.134.$$

So we can see, we can minimizing the objective function simply by scaling.

An Example

Consider a extreme case, $\bar{w}^T x_i = +\infty$. Then

$$\log(1 + e^{-\infty}) = 0.$$

note that 0 cannnot be actually attained.
As our goal is to minimize the total loss, so the optimal objective function value can be very small simply by scaling of the hypothesis.
Logistic loss is sensitive to SCALING

Drawbacks

One may argue that as the sample problem is separable, it is OK to choose such 'big' w and b. However, sample data being separable doesn't necessarily implies that the whole set is separable.
Consider an example that some actual data violates the hypothesis.
If $w^T x + b = 1000$, and there exists a $y_i < 0$ in actual data, then

$$\log(1 + e^{1000}) \approx +\infty,$$

which means the error of such logistic loss function with big 'w' and 'b' can be very huge.

Drawbacks

- There might be huge difference between expected error and empirical error.
- Sample separable doesn't imply actual data separable.

How can we modify logistic loss function to eliminate its drawbacks?

Logistic loss and regularization

Regularized logistic regression

$$\min_{w:\|w\|\leq B} f(w) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i)})$$

Or

$$\min_{w} f(w) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y_i(w^T x_i)}) + \lambda\|w\|^2$$

Large Margin Classification

Assume that w is normalized so that $\|w\|_2 = 1$.
We do not only want to separate points, but we want to separate them **with a margin** :

$$w^T x_i + b \geq \gamma \text{ if } y_i = 1$$
$$w^T x_i + b \leq -\gamma \text{ if } y_i = -1$$

Alternately, we can express both of the statements above as the single requirement

$$y_i \left(w^T x_i + b \right) \geq \gamma \quad i = 1, \ldots, m$$

Large Margin Classification

We want the largest margin (why?)

$$\begin{aligned} &\underset{w,b,\gamma}{\text{maximize}} \ \gamma \\ &\text{subject to} \ \ y_i \left(w^T x_i + b \right) \geq \gamma \quad i = 1, \ldots, m \\ &\qquad\qquad \|w\|_2 = 1 \end{aligned}$$

Note that as stated, the problem above is not convex. Specifically, because the equality constraint $\|w\|_2 = 1$ is not affine.

Large Margin Classification

We will show that this can be rewritten as

$$\begin{aligned} &\underset{\tilde{w},\tilde{b}}{\text{minimize}} \quad \|\tilde{w}\|_2 \\ &\text{subject to} \quad y_i\left(\tilde{w}^T x_i + \tilde{b}\right) \geq 1 \quad i = 1, \ldots, m \end{aligned}$$

- This is a convex problem.
- Note that there is no loss function, because we assume that we achieve zero loss.
- Moreover we use a different loss function - a margin loss and assume it is equal to zero.
- We will generalize this later.

Rademacher Complexity

Given a hypotheses set $\mathcal{H}$, The Rademacher complexity of the function class $\mathcal{H}$ for sample size m is:

$$\mathcal{RC}_m(\mathcal{H}) = \mathbb{E}_x \mathbb{E}_{\sigma \sim unif\{\pm 1\}^n} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i h(x_i) \right| \right].$$

Some features:

- If $\|w\| \to \infty$, then $\mathcal{RC}_m(\mathcal{H}) \to \infty$.
- Sensitive to scaling.
- Connected with VC-dimension.
- $\mathbb{E}_x$ can be replaced with the worst case to bound $\mathcal{R}_m(\mathcal{H})$.

Rademacher Complexity

Recall $\mathcal{R}(h) = \mathbb{E}(loss(h(x), y))$, $\hat{\mathcal{R}}(h) = \frac{1}{m}\sum loss(h(x_i), y_i)$. And consider $\mathcal{H}_B = \{w : x \to w^T x | \|w\|_2 \leq B\}$, so we have,

$$\mathcal{RC}(\hat{h}) \leq \mathcal{R}(h^*) + \tilde{\mathcal{O}}[L \cdot \mathcal{RC}_m^2(\mathcal{H}) + \sqrt{L \cdot \mathcal{R}(h^*)}\mathcal{RC}_m(\mathcal{H})],$$

where L is the Lipschitz constant of loss function.
We can bound Rademacher complexity

$$\mathcal{RC}_m(\mathcal{H}_B) \leq XB\sqrt{\frac{2}{m}},$$

where $X : \{\|x\|_2 \leq X, \forall x \in \mathcal{X}\}$ and $B : \{\|w\|_2 \leq B, \forall w \in \mathcal{H}_B\|\}$

γ-fat shattering

Now instead of VC-dimension (aka shattering), we will have a "fat-shattering" dimension.

Definition

The points $x_1, \ldots, x_m$ are γ-fat shattered by the hypothesis class $\mathcal{H} \in \mathbb{R}^{\mathcal{X}}$ if for all possible labelings $y_1, \ldots, y_m$, there exists a predictor $h \in \mathcal{H}$ such that every labeling is possible with a margin of γ. That is $y_i \left(h(x_i)\right) \geq \gamma$.

Definition

The γ-fat shattering (or "fat shattering at scale γ") dimension of a hypothesis class $\mathcal{H} \in \mathbb{R}^{\mathcal{X}}$ is the largest number of points m such that there exists some set of points $x_1, \ldots, x_m$ that are γ-fat shattered.

Example of γ-fat shattering for the case of a restricted classifier set.

$$\mathcal{H}_{w;B} = \left\{ x \mapsto w^T x \mid \|w\|_2 \leq B \right\}$$
$$\mathcal{X} = \{ x \mid \|x\|_2 \leq X \}.$$

For this setting, the γ-fat shattering dimension $D_\gamma \propto \frac{B^2 X^2}{\gamma^2}$. Note that these parameters $(B,\ X,\ \gamma)$ must scale together.

- Can 1 point be γ-fat shattered with $\gamma = 0.9$? Yes.
- Can 2 points be γ-fat shattered with $\gamma = 0.9$?
- Can 3 points be γ-fat shattered with $\gamma = 0.9$?

Width of the margin over the radius of the data, geometrically $\frac{BX}{\gamma}$. The ratio is what matters.
The relative margin is $\frac{\gamma}{BX}$.

Deviation bounds involving the fat-shattering dimension

For an L-Lipschitz loss function, the γ-fat shattering dimension controls the capacity/complexity.

$$\mathcal{H}_{w;B} = \left\{ x \mapsto w^T x \mid \|w\|_2 \leq B \right\}$$
$$\mathcal{X} = \left\{ x \mid \|x\|_2 \leq X \right\}.$$

Lemma

For any loss that is L-Lipschitz, with probability at least $1 - \delta$

$$\left| R(h) - \hat{R}(h) \right| \leq 2 \cdot L \sqrt{\frac{B^2 X^2}{m}} + s_{max} \sqrt{\frac{\log \frac{2}{\delta}}{m}}.$$

Regularized empirical risk minimization

$$\mathcal{H}_{w;B} = \left\{ x \mapsto w^T x \mid \|w\|_2 \leq B \right\}$$
$$\mathcal{X} = \{ x \mid \|x\|_2 \leq X \}.$$

For any loss that is L-Lipschitz,

$$R(h) \leq \hat{R}(h) + 2 \cdot L \sqrt{\frac{B^2 X^2}{m}} + s_{\max} \sqrt{\frac{\log \frac{2}{\delta}}{m}}.$$

R, δ and L are constants beyond our control (more or less). Assume m is also fixed. But B can be controlled

$$R\left(\hat{h}\right) \leq \min_{\|w\| \leq B} \left[\hat{R}(h) + 2 \cdot L \sqrt{\frac{B^2 X^2}{m}} + s_{\max} \sqrt{\frac{\log \frac{2}{\delta}}{m}} \right]$$