MATH 565: Lecture 8 (02/05/2026)

Today: * SGD for SVM
* Logistic regression loss
* coordinate descent (CD)

Recall

$$\nabla J_{H\text{-}SVM} = -y_i \bar{x}_i \, \delta\left(1 - y_i(\bar{w}^T \bar{x}_i) > 0\right) + \lambda \bar{w} \qquad \text{(Hinge-SVM)}$$

indicator function

$$\nabla J_{L_2\text{-}SVM} = -y_i \bar{x}_i \max\{0, 1 - y_i(\bar{w}^T \bar{x}_i)\} + \lambda \bar{w} \qquad (L_2\text{-SVM})$$

SGD for H-SVM

$$\bar{w} \leftarrow \bar{w}(1-\alpha\lambda) + \alpha \sum_{\cancel{i=1}\; i \in S}^{\cancel{n}} y_i \bar{x}_i \, \delta\left(1 - y_i(\bar{w}^T \bar{x}_i) > 0\right)$$

Let $S^+ = \{ i \in S \mid y_i(\bar{w}^T \bar{x}_i) < 1 \}$ → subset of indices in S for which $\delta(\cdot) = 1$ above

$y_i(\bar{w}^T \bar{x}_i) < 0 \implies i$: misclassified point/instance

$y_i(\bar{w}^T \bar{x}_i) \in (0,1) \implies i$: correctly classified instance, but lies close to decision boundary.

When $y_i(\bar{w}^T \bar{x}_i) \geq 1$, i is correctly classified and well-separated
$\implies$ does not contribute to J.

SGD for Hinge-SVM

$$\bar{w} \leftarrow \bar{w}(1-\alpha\lambda) + \sum_{i \in S^+} \alpha y_i \bar{x}_i \quad \longrightarrow \text{ primal SVM algorithm}$$

proposed by Hinton in 1989!

↳ before VC dimension and other details were proposed by Vapnik and coauthors later on...

Logistic Regression Loss

Note that the hinge loss function, while convex, is not smooth — there is a sharp "hinge" at the value of 1 (hence the name). The logistic regression loss can be considered as a smooth version of the hinge loss.

$$J_{LR} = \sum_{i=1}^{n} \log\left(1+e^{-y_i(\bar{w}^T\bar{x}_i)}\right) + \frac{1}{2}\|\bar{w}\|^2$$

Consider $L(z) = \log(1+e^{-z})$ with $z = yf(\bar{x})$ as the prediction

e.g., $z_i = y_i(\bar{w}^T\bar{x}_i)$

$$= \log\left(e^{-z}(1+e^{z})\right)$$

$$= -z + \underbrace{\log(1+e^{z})}_{\to 0 \text{ as } z \to -\infty}$$

largely misclassified instances give huge negative values

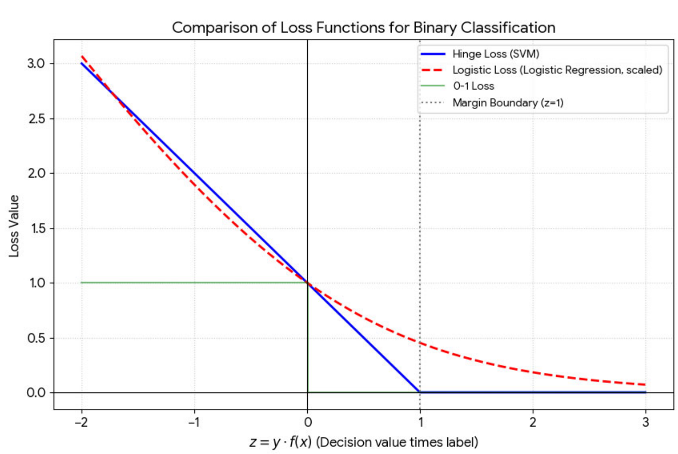
$\Rightarrow$ For largely misclassified instances, J_{LR} increases linearly as $|\bar{w}^T\bar{x}_i|$ increases. For such instances

$$J_{H\text{-}SVM} - J_{LR} \approx 1$$

recall that the hinge loss function is $1-z$ for $z<1$...

$\Rightarrow$ H-SVM and LR-SVM treat grossly misclassified instances similarly.

But $J_{LR} > 0$ (ignoring $\|\bar{w}\|^2$ term) for all instances.



J_{LR} is differentiable while $J_{H\text{-}SVM}$ is not.

In fact, the fit term in the J_{LR} loss function turns out to be strictly convex (on its own)!

Lemma 8 J_{LR} without the $\frac{1}{2}\|\bar{w}\|^2$ regularizer term is strictly convex.

But we usually add $\frac{1}{2}\|\bar{w}\|^2$ still, as this extra term encourages sparsity. We could instead use $\sum_{i=1}^{n} |w_i|$ as an L_1-regularity term. But then again, $|w_i|$ is not smooth either...

We finish with the details of gradient descent using the logistic regression loss function for SVM.

$$\nabla J_{LR} = -\sum_{i=1}^{n} \frac{y_i \bar{x}_i}{\left[1+e^{-y_i(\bar{w}^T\bar{x}_i)}\right]} + \lambda \bar{w}$$

Hence, the SGD update is given as follows.

$$\bar{w} \leftarrow \bar{w}(1-\alpha\lambda) + \sum_{i \in S} \frac{\alpha\, y_i \bar{x}_i}{\left[1+e^{-y_i(\bar{w}^T\bar{x}_i)}\right]}.$$

Coordinate Descent (CD)

Recall the gradient descent update: $\bar{w} \leftarrow \bar{w} - \alpha \nabla J$.

In coordinate descent (CD), we optimize one coordinate at a time.

$$\bar{w} = \arg\min_{\bar{w}} \{ J(\bar{w}) \mid \text{only } w_i \text{ varies} \}$$

* only one variable to handle — can be (much) easier.
* can use line search if not able to solve exactly.

Cycle through all $i = 1, \dots, d$.
if no w_i changes, STOP. → in a full cycle

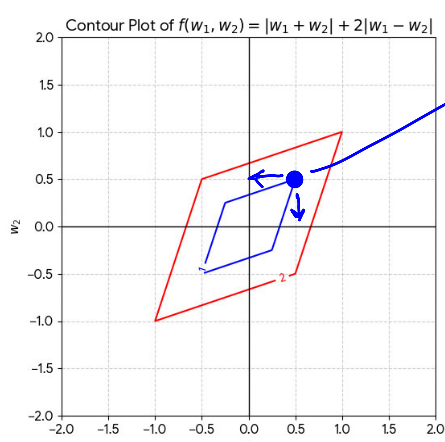
* If J is convex and differentiable, then the converged solution is optimal.
* But if J is not differentiable, this guarantee does not hold, even when it may be convex.

Consider $J(\bar{w}) = |w_1 + w_2| + \alpha |w_1 - w_2|$, $\alpha > 1$. → can show that J is convex.

J is minimal at $(0,0)$.

But if we're at $(1,1)$, neither coordinate will decrease J.

two contours of $J=1$, $J=2$.



Still, we can give a characterization of a fairly general class of loss functions on which CD is well-behaved.

Lemma 9 Let $J(\bar{w}) = G(\bar{w}) + \underbrace{\sum_{i=1}^{d_1} H_i(w_i)}_{}$ where $G(\bar{w})$ is convex and differentiable, while $H_i(w_i)$ are convex but may not be differentiable. Then, CD converges to the global minimum of J.

single variable functions!

The L_1-regularizer has this structure: $H_i = |w_i|$, giving $\sum_{i=1}^{d_1} |w_i|$ as the regularizer term.

But in some cases, variable transformations can help even if J is not in this form.

e.g., $J(\bar{w}) = G(\bar{w}) + |w_1+w_2| + \alpha|w_1 - w_2|$, $\alpha > 1$. Here,

convex, differentiable

we can use $u_1 = \frac{w_1+w_2}{2}$ and $u_2 = \frac{w_1 - w_2}{2}$ to get

$w_1 = u_1 + u_2$ and $w_2 = u_1 - u_2$, giving

$$J(\bar{u}) = G(u_1+u_2, u_1-u_2) + 2|u_1| + 2\alpha|u_2|,$$

which has structure specified in Lemma 9.

CD will work well on this version of J.

Linear Regression with Coordinate Descent

To understand CD better, we apply it to linear regression (without regularization)

$$J(\bar{w}) = \frac{1}{2}\|D\bar{w}-\bar{y}\|^2 = \frac{1}{2}\sum_{i=1}^{n}(\bar{w}^T x_i - y_i)^2$$

$$= \frac{1}{2}\sum_{i=1}^{n}\Big(\underbrace{\sum_{j=1}^{d} w_j x_{ij} - y_i}_{w_k x_{ik} + \sum_{j\neq k} w_j x_{ij} - y_i}\Big)^2$$

we consider CD for w_k

$$\frac{\partial J}{\partial w_k} = \sum_{i=1}^{n}\Big(w_k x_{ik} + \sum_{j\neq k} w_j x_{ij} - y_i\Big)x_{ik} = 0$$

first order optimality

$$\Rightarrow w_k = \frac{-\sum_{i=1}^{n}\Big(\sum_{j\neq k} w_j x_{ij} - y_i\Big)x_{ik}}{\sum_{i=1}^{n} x_{ik}^2}$$

$\Big(y_i - \sum_{j\neq k} w_j x_{ij}\Big)$

term not included

With $\bar{r} = \bar{y} - D^T\bar{w} = \bar{y} - \sum_{j\neq k}\bar{d}_j w_j - w_k \bar{d}_k$

vector of residuals

$\bar{d}_j$: j^{th} column of D.

The update step is gives as follows.

$$w_k^{new} = \frac{\bar{d}_k^T(\bar{r} + w_k^{old}\bar{d}_k)}{\|\bar{d}_k\|^2}$$

$$D = \begin{array}{c|c} & 1 \; 2 \; \cdots \; k \; \cdots \; d \\ \begin{matrix}1\\2\\\vdots\\\vdots\\n\end{matrix} & \begin{bmatrix} & & \boxed{\begin{matrix}x_{1k}\\x_{2k}\\\vdots\\x_{ik}\\\vdots\\x_{nk}\end{matrix}} & \end{bmatrix}\end{array}$$

$$= w_k^{old} + \frac{\bar{d}_k^T\bar{r}}{\|\bar{d}_k\|^2}$$

we assume the trivial case of $\bar{d}_k = \bar{0}$ does not occur.

If data columns are normalized, $\|\bar{d}_k\|^2 = 1$, and we get $w_k^{new} = w_k^{old} + \bar{d}_k^T\bar{r}$.

More generally, we update

$$w_k^{new} \leftarrow w_k^{old} + \bar{d}_k^T \bar{r} \longrightarrow \text{extremely efficient!}$$

We also update the residuals $\bar{r}$ as follows:

$$\bar{r} \leftarrow \bar{r} - \bar{d}_k (\Delta w_k) \longrightarrow w_k^{new} - w_k^{old}$$