

ECE 532 - lecture 24 - duality

①

the dual of an optimization problem can mean many things.

e.g. "Lagrange dual". In the context of classification problems, it means an alternate formulation that is equivalent (same result) but perhaps easier to compute/solve.

Consider a standard LS problem with L2 regularization.

In the classification setting, we have features x_i with labels y_i .

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad y \in \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m. \quad \text{We seek weights } w \in \mathbb{R}^n$$

so that we minimize:

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2$$

$$= \min_w \sum_{i=1}^m (1 - y_i x_i^T w)^2 + \lambda \|w\|^2 \quad (\text{when } y_i \text{ labels are } \pm 1)$$

* Claim: the optimal w^* is of the form

$$w^* = X^T \alpha = \sum_{j=1}^m \alpha_j x_j, \quad \text{i.e. a linear combination of the features.}$$

Prof.. Suppose $w^* = \sum_{j=1}^m \alpha_j x_j + x^\perp$ (2)

where $x^\perp \in \text{span}\{x_1, \dots, x_m\}^\perp$, our claim is that $x^\perp = 0$,

substitute this formula for w into the optimization problem:

$$\min_{\alpha, x^\perp} \sum_{i=1}^m \left(1 - y_i x_i^\top \left(\sum_{j=1}^m \alpha_j x_j + x^\perp\right)\right)^2 + \lambda \left\| \sum_{j=1}^m \alpha_j x_j + x^\perp \right\|^2$$

$$= \min_{\alpha, x^\perp} \sum_{i=1}^m \left(1 - y_i x_i^\top \sum_{j=1}^m \alpha_j x_j\right)^2 + \lambda \left\| \sum_{j=1}^m \alpha_j x_j \right\|^2 + \underbrace{\lambda \|x^\perp\|^2}_{\text{can make this zero.}}$$

where we used the fact that $x_i^\top x^\perp = 0$ for all i .

So clearly, we should pick $x^\perp = 0$ to minimize.

or, in matrix form:

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2 \quad \text{let } w = X^\top \alpha.$$

$$= \min_\alpha \|XX^\top \alpha - y\|^2 + \lambda \|X^\top \alpha\|^2$$

$$= \min_\alpha \|XX^\top \alpha - y\|^2 + \lambda \alpha^\top XX^\top \alpha \quad \text{let } XX^\top = K.$$

$$= \min_\alpha \|K\alpha - y\|^2 + \lambda \alpha^\top K\alpha \quad (K_{ij} = x_i^\top x_j).$$

This is a regularized least-squares problem! But variable is $\alpha \in \mathbb{R}^m$ instead of $w \in \mathbb{R}^n$.

(3)

$$\min_{\alpha} \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha$$

This is called the dual formulation.

We can solve by taking the derivative:

$$\begin{aligned} \frac{\partial}{\partial \alpha} \{ \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha \} \\ = 2K^T(K\alpha - y) + 2\lambda K \alpha = 0. \end{aligned}$$

\nwarrow note $K^T = K$.

$$\Rightarrow Ky = K \underbrace{(K + \lambda I)}_{\text{always invertible.}} \alpha$$

$$\text{Solution is } \alpha = (K + \lambda I)^{-1}y.$$

This makes sense, because if we reconstruct w from α :

$$\begin{aligned} w = X^T \alpha &= X^T (K + \lambda I)^{-1}y = X^T (X X^T + \lambda I)^{-1}y \\ &= (X^T X + \lambda I)^{-1} X^T y. \end{aligned}$$

i.e. we recover the solution we were expecting.

This gives us the option of solving our problem by solving

$$(i) \min_w \|Xw - y\|^2 + \lambda \|w\|^2 \quad (\text{the } \underline{\text{primal}})$$

$$\text{or } (ii) \min_{\alpha} \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha \quad (\text{the } \underline{\text{dual}})$$

(with α and w related via $w = X^T \alpha$.)

(4)

If we are using kernels, then the LS problem becomes:

$$\min_w \|\Phi w - y\|^2 + \lambda \|w\|^2$$

This is an optimization in $w \in \mathbb{R}^N$.

where $\Phi \in \mathbb{R}^{m \times N}$ is the vector of features

$$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_m)^T \end{bmatrix}.$$

But the (equivalent) dual is an optimization in $\alpha \in \mathbb{R}^m$

$$\min_{\alpha} \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha$$

and we only need to know the kernel matrix $K \in \mathbb{R}^{m \times m}$ to solve it!

The same argument holds for kernelized SVM and regular SVM.

Consider the same problem setup (classification) but with hinge loss (SVM):

$$\min_w \sum_{i=1}^m (1 - y_i x_i^T w)_+ + \lambda \|w\|^2.$$

Once again, we can let $w = \sum_{j=1}^m \alpha_j x_j + x^+$

and we conclude that $x^+ = 0$ at optimality.

Substituting, we obtain:

$$\min_{w \in \mathbb{R}^m} \sum_{i=1}^m \left((1 - y_i x_i^T \sum_{j=1}^m \alpha_j x_j)_+ + \lambda \left\| \sum_{j=1}^m \alpha_j x_j \right\|^2 \right)$$

(cont'd)

$$\begin{aligned}
 &= \min_{\alpha} \sum_{i=1}^m \left(1 - y_i \sum_{j=1}^m \alpha_j x_i^T x_j \right)_+ + \lambda \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j x_i^T x_j \\
 &= \min_{\alpha} \sum_{i=1}^m \left(1 - y_i \sum_{j=1}^m K_{ij} \alpha_j \right)_+ + \lambda \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K_{ij} \\
 &= \min_{\alpha} \mathbb{1}^T \left(\mathbb{1} - y \circ (K\alpha) \right)_+ + \lambda \alpha^T K \alpha
 \end{aligned}$$

So the dual of the L_2 -regularized SVM in $w \in \mathbb{R}^n$ is also an L_2 -regularized SVM, but this time in $\alpha \in \mathbb{R}^m$.

The steps are as follows

- 1) solve for $\hat{\alpha}$ via the dual.
- 2) if desired, compute $\hat{w} = X^T \hat{\alpha}$ (or $\Phi^T \hat{\alpha}$).
- 3) to classify a new example \tilde{x} , compute
 $\hat{y} = \text{sign}(\tilde{x}^T \hat{w}) = \text{sign}(\tilde{x}^T X^T \hat{\alpha}) = \text{sign}\left(\sum_{i=1}^m k(\tilde{x}, x_i) \hat{\alpha}_i\right)$
 or, in the kernelized case:
 $\hat{y} = \text{sign}(\phi(\tilde{x})^T \hat{w}) = \text{sign}(\phi(\tilde{x})^T \Phi^T \hat{\alpha}) = \text{sign}\left(\sum_{i=1}^m k(\tilde{x}, x_i) \hat{\alpha}_i\right)$

No need to compute \hat{w} to classify new examples.

(6)

the kernel function $k(x, y)$ or in this case the kernel matrix K where $K_{ij} = k(x_i, x_j)$ intuitively measures the similarity between two examples.

Roughly speaking, the "Kernel trick" that allows us to bypass the representation $\phi(x)$ and have efficient computation is possible whenever $K \succeq 0$ (positive semidefinite).

Examples include:

$$\text{"linear kernel": } k(x, y) = x^T y$$

$$\text{"polynomial kernel": } k(x, y) = (x^T y + 1)^d$$

$$\text{"Gaussian kernel": } k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

$$\text{"Laplacian kernel": } k(x, y) = \exp(-\alpha \|x - y\|)$$

Also, a covariance matrix can be used as a kernel if we know something about the statistical properties of the features.