MULTI-TASK AND TRANSFER LEARNING

Consider a supervised learning setting where P_{XY} denotes the joint distribution on (X, Y).

Consider distributions $P_{XY}^{(1)}, \ldots, P_{XY}^{(N)}$ and training data

$$(X_j^{(i)}, y_j^{(i)}) \sim P_{XY}^{(i)}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, N.$$

Furthermore, suppose the distribution $P_{XY}^{(i)}$ are related, so that the optimal classifiers / regression functions are similar.

Multi-task learning: Learn the predictors of the N tasks simultaneously

Transfer Learning: Learn the predictor of a new task $P_{XY}^{(0)}$ from which only unlabeled examples are available, $X_1^{(0)}, \ldots, X_{n_0}^{(0)} \sim P_X^{(0)}$ (X-marginal distribution of $P_{XY}^{(0)}$). [Note: some papers on transfer learning also assume a few labels from P_{XY}^{T} are available]

Multitask Learning

Let's focus on the linear setting.

Let $w^{(i)} \in \mathbb{R}^d$ denote the predictor associated with task i.

$$W := \begin{bmatrix} w^{(1)} & \cdots & w^{(N)} \end{bmatrix} = \begin{bmatrix} w_1^T \\ \vdots \\ w_d^T \end{bmatrix} \qquad (d \times N)$$

The basic idea is to solve

$$\min_W \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell\left(y_j^{(i)}, \langle w^{(i)}, x_j^{(i)} \rangle\right) + \lambda R(W)$$

where ℓ is a loss (e.g., squared error, hinge, logistic) and R is a penalty/regularization term that encourages $w^{(1)}, \ldots, w^{(N)}$ to be similar. Can you suggest a good R?

- shared mean: $R(W) = \sum_{i=1}^{N} \left\| W^{(i)} - \frac{1}{N} \sum_{k=1}^{N} W^{(k)} \right\|_2^2$

  Note: this can be kernelized

- low rank: $R(W) = \|W\|_*$ (nuclear norm = sum of singular values, a convex upper bound on the rank)

- group lasso: $R(W) = \sum_{\ell=1}^{d} \|W_\ell\|_2$

Aside: Group Lasso

Recall the lasso is a penalty that promotes sparsity and is useful for feature selection.

The group lasso is useful for "group feature selection."

Consider a prediction problem (classification or regression) where the features can be naturally grouped.

Examples
- classification of brain images, groups correspond to anatomical units (e.g., hippocampus, visual cortex)
- ?

Let $G_1, \ldots, G_m$ be a partition of $\{1, 2, \ldots, d\}$.

Then $G_r \cap G_s = \emptyset$, $\bigcup_{r=1}^{m} G_r = \{1, \ldots, d\}$.

Let w_G denote the vector w restricted to features in G, e.g.

$$w = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}, \quad G = \{2, 5\} \implies w_G = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

The group lasso penalty is $\sum_{r=1}^{m} \|w_{G_r}\|_2$.

This can be thought of as the l_1 norm applied to $(\|w_{G_1}\|, \|w_{G_2}\|, \ldots, \|w_{G_m}\|)$.

Therefore, the group lasso penalty leads to

$\|w_{G_r}\| = 0$ for most groups $\implies$ group feature selection.

Example: Linear regression w/ group feature selection

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{r=1}^{m} \|w_{G_r}\|_2$$

(A) What are the groups in multitask learning?

Further Aside: Multiclass SVM

How would you generalize the SVM from binary to multiclass?

Let's consider a linear SVM. One way to define the svm in the multiclass case is

$$f(x) = \arg\max_{k=1,\dots,K} \langle w_k, x \rangle$$

where $w_k \in \mathbb{R}^d$ is associated w/ class k, and solves

$$\min_{w_1,\dots,w_K} \frac{1}{2}\sum_{k=1}^{K} \|w_k\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \langle w_{y_i} - w_k, x_i \rangle \geq 1 - \xi_i \quad \forall k \neq y_i, \ \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

(B) How could we jointly train a multiclass SVM to also select relevant features?

Transfer Learning

The first thing to notice is that this is a different kind of supervised learning problem.

Normal classification:

Input: feature vector x

Output: Label y

Transfer Learning:

Input: Unlabeled data $S^{(0)} = \{x_1^{(0)}, \ldots, x_{n_0}^{(0)}\}$

Output: Classifier

One representation of the input-output relationship is $g(S^{(0)})$. Since this is a classifier, it makes sense to write $g(S^{(0)})(x_j^{(0)})$, which is the predicted label.

Instead, let's use the equivalent expression

$$f(S^{(0)}, x_j^{(0)}) := g(S^{(0)}, x_j^{(0)})$$

viewed as a function on $\mathcal{D} \times \mathbb{R}^d$, where

$\mathcal{D}$ = set of all unlabeled datasets.

Now we can use the RKHS framework. Thus, suppose

$k =$ a PSD kernel on $\mathcal{D} \times \mathbb{R}^d$, and let $\mathcal{H}$ denote its RKHS.

(more on this below)

Then f is the solution of

$$\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} l\left(y_i, f\left(S^{(i)}, x_j^{(i)}\right)\right) + \lambda \|f\|^2$$

where, e.g., l is the hinge loss, $l(y,t) = \max(0, 1-yt)$.

By the representer theorem, f has the form

$$f(S,x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \alpha_j^{(i)} k\left(\left(S^{(i)}, x_j^{(i)}\right), (S,x)\right)$$

and $\{\alpha_j^{(i)}\}$ can be obtained using existing optimization strategies (e.g., for hinge loss, this is equivalent to a certain cost-sensitive SVM).

Defining a kernel on $\mathcal{D} \times \mathbb{R}^d$

If $k_{\mathcal{D}}$ is a kernel on $\mathcal{D}$ and k_x is a kernel on $\mathbb{R}^d$, then

$$k((S,x),(S',x')) = k_{\mathcal{D}}(S,S') \cdot k_x(x,x')$$

is a kernel on $\mathcal{D} \times \mathbb{R}^d$.

- kernel on $\mathbb{R}^d$: e.g., Gaussian kernel

$$k_X(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$$

- kernel on $\mathbb{D}$: define

$$\Psi(S) = \Psi(\{x_1, \ldots, x_n\})$$

$$= \frac{1}{n}\sum k_x(\cdot, x_j)$$

Then the kernel can be defined to equal

$$k_{\mathbb{D}}(S, S') = \exp\left(-\frac{1}{2\sigma^2}\|\Psi(S) - \Psi(S')\|^2\right)$$

↑ norm in RKHS assoc. w/ k_X - can be evaluated using reproducing property.

Addition details: Blanchard, Lee, Scott, "Generalizing from Several Related Tasks to A New Unlabeled Sample."

Key A. The groups are the entries of W in the same row.

B. Group lasso penalty where groups correspond to features - same structure as multi-task learning.