What is semi-supervised learning ?

- In many practical learning domains, there is a large supply of unlabeled data but limited labeled data, which can be expensive to generate
  - text processing, video-indexing, bioinformatics
- Semi-supervised Learning: learning from a combination of both labeled and unlabeled data

Comparing

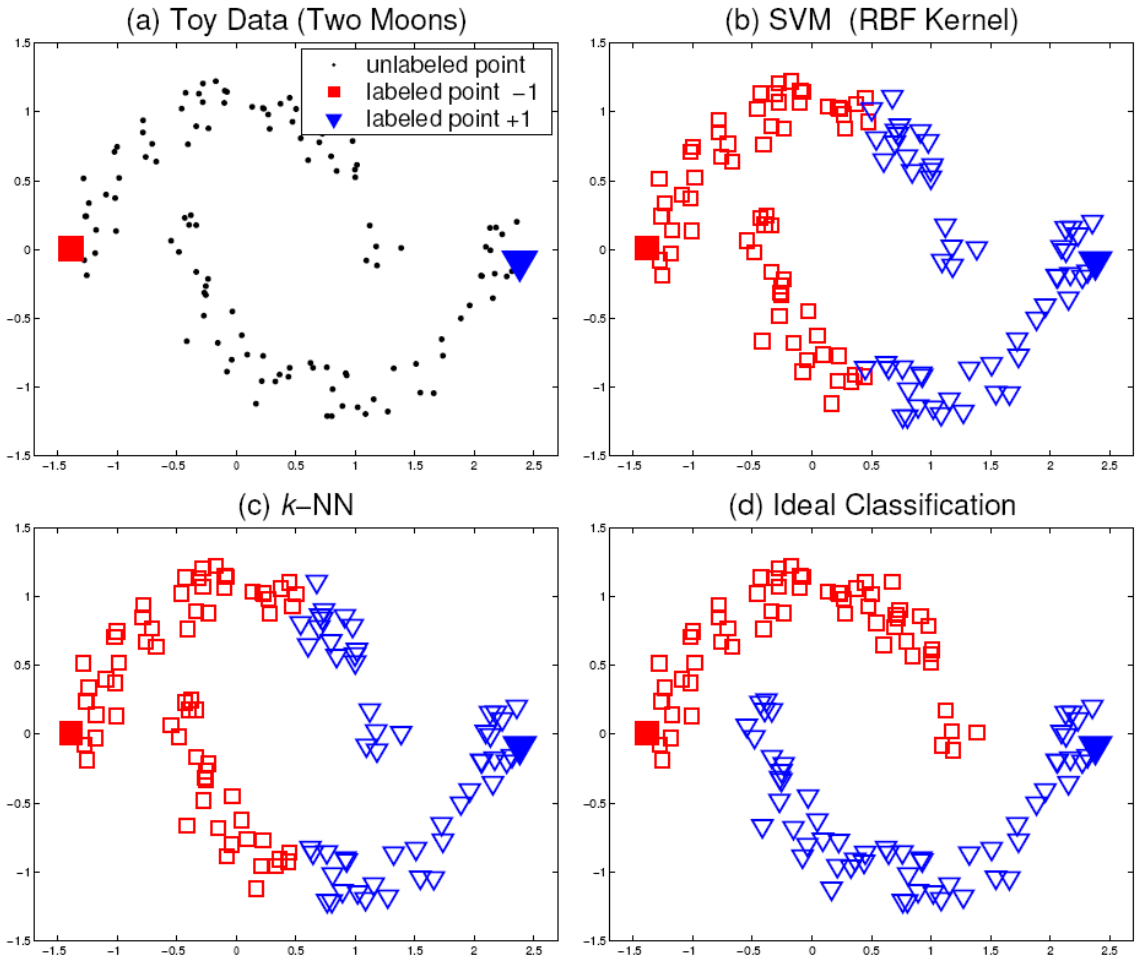
- Supervised learning algorithms require enough labeled training data to learn reasonably accurate classifiers.
- Unsupervised learning methods are employed to discover structure in unlabeled data
- Semi-supervised learning allows taking advantage of the strengths of both

Why should it be useful ?

Unlabeled data can help in two different ways

- Identify data structure
  Find a meaningful representation of complicated high dimensional data through a first unsupervised learning step.
- Cluster assumption
  which can be stated in two equivalent ways:
  - Two points which can be connected by a high density path (i.e. in the same cluster) are likely to be of the same label.
  - Decision boundary should lie in a low density region.

A Toy Dataset (Two Moons)



Learning from Examples

- Input space $\mathcal{X}$, and output space $\mathcal{Y} = \{1, -1\}$.
- Training set $S = \{z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)\}$ in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from some unknown distribution.
- Classifier $f : \mathcal{X} \to \mathcal{Y}$.

Transductive Setting

- Input space $\mathcal{X} = \{x_1, \ldots, x_n\}$, and output space $\mathcal{Y} = \{1, -1\}$.
- Training set $S = \{z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)\}$.
- Classifier $f : \mathcal{X} \to \mathcal{Y}$.

Intuition about classification: Manifold

- **Local consistency**. Nearby points are likely to have the same label.
- **Global consistency**. Points on the same structure (typically referred to as a cluster or manifold) are likely to have the same label.

Algorithm

1. Form the affinity matrix W defined by $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2. Construct the matrix $S = D^{-1/2}WD^{-1/2}$ in which D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W.
3. Iterate $f(t+1) = \alpha S f(t) + (1-\alpha)y$ until convergence, where α is a parameter in $(0,1)$.
4. Let f^* denote the limit of the sequence $\{f(t)\}$. Label each point x_i as $y_i =$ sgn(f_i).

Convergence

Theorem. *The sequence* $\{f(t)\}$ *converges to* $f^* = \beta(I - \alpha S)^{-1}y$, *where* $\beta = 1 - \alpha$.

Proof. Suppose $F(0) = Y$. By the iteration equation, we have

$$f(t) = (\alpha S)^{t-1}Y + (1-\alpha)\sum_{i=0}^{t-1}(\alpha S)^i Y. \tag{1}$$

Since $0 < \alpha < 1$ and the eigenvalues of S in $[-1, 1]$,

$$\lim_{t\to\infty}(\alpha S)^{t-1} = 0, \text{ and } \lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha S)^i = (I - \alpha S)^{-1}. \tag{2}$$

Regularization Framework

Cost function

$$\mathcal{Q}(f) = \frac{1}{2}\left[\sum_{i,j=1}^{n} W_{ij}\left(\frac{1}{\sqrt{D_{ii}}}f_i - \frac{1}{\sqrt{D_{jj}}}f_j\right)^2 + \mu\sum_{i=1}^{n}\left(f_i - y_i\right)^2\right]$$

- **Smoothness term.** Measure the changes between nearby points.
- **Fitting term.** Measure the changes from the initial label assignments.

Regularization Framework

Theorem. $f^* = \arg\min_{f \in \mathcal{F}} \mathcal{Q}(f)$.

Proof. Differentiating $\mathcal{Q}(f)$ with respect to f, we have

$$\left.\frac{\partial \mathcal{Q}}{\partial f}\right|_{f=f^*} = f^* - Sf^* + \mu(f^* - y) = 0, \tag{1}$$

which can be transformed into

$$f^* - \frac{1}{1+\mu} Sf^* - \frac{\mu}{1+\mu} y = 0. \tag{2}$$

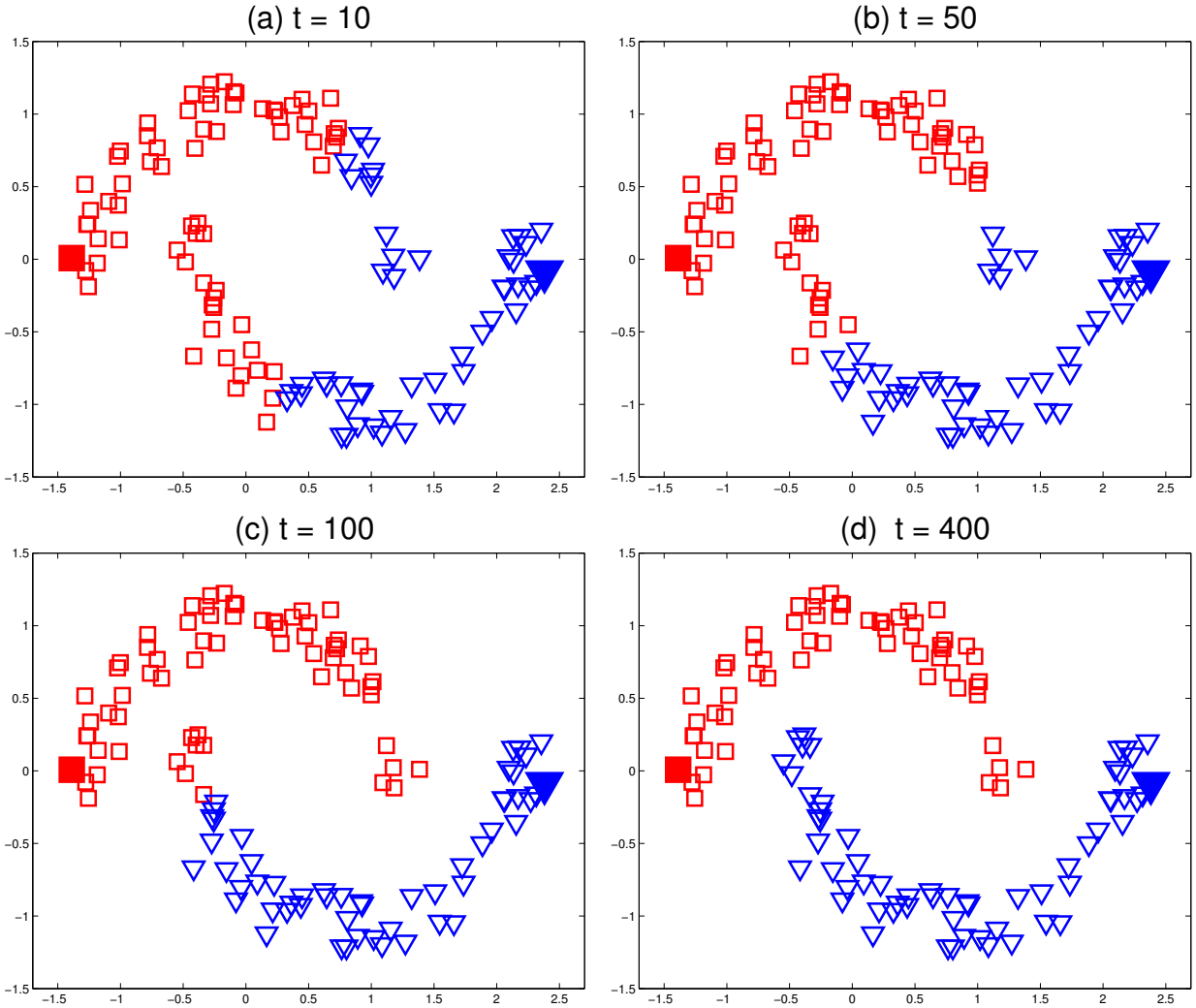
Let $\alpha = 1/(1+\mu)$ and $\beta = \mu/(1+\mu)$. Then

$$(I - \alpha S) f^* = \beta y. \tag{3}$$

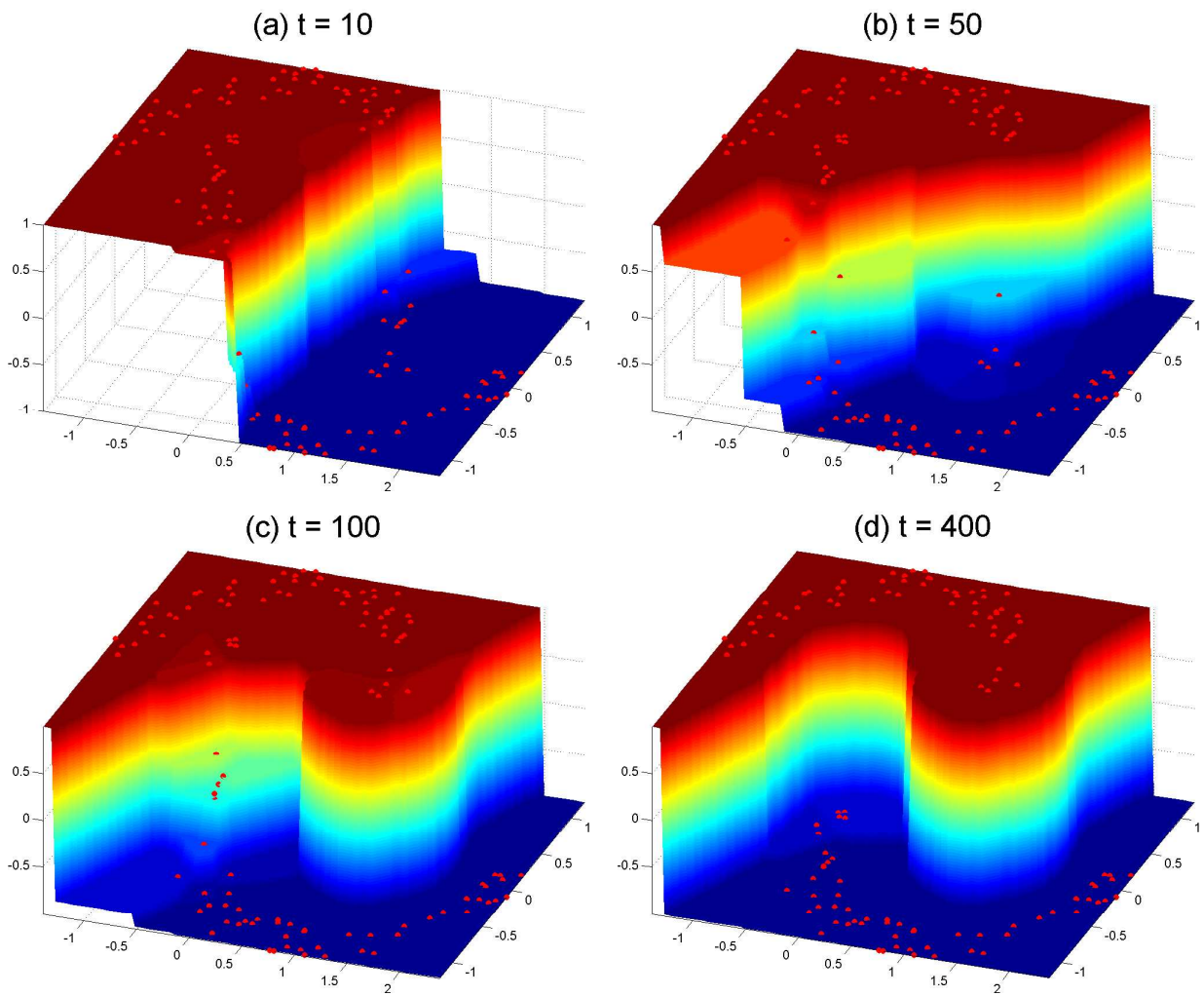
Two Variants

- Substitute $P = D^{-1}W$ for S in the iteration equation. Then $f^* = (I - \alpha P)^{-1}y$.
- Replace S with P^T, the transpose of P. Then $f^* = (I - \alpha P^T)^{-1}y$, which is equivalent to $f^* = (D - \alpha W)^{-1}y$.

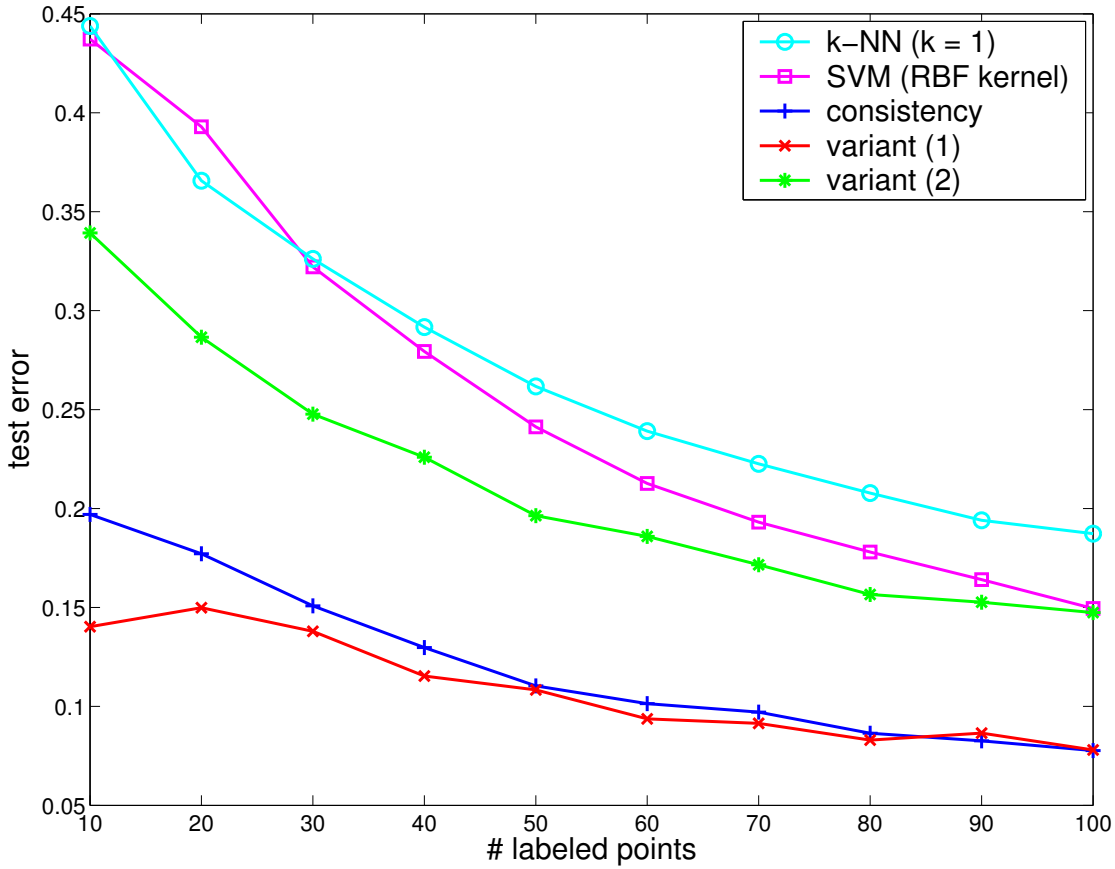
Toy Problem



Toy Problem

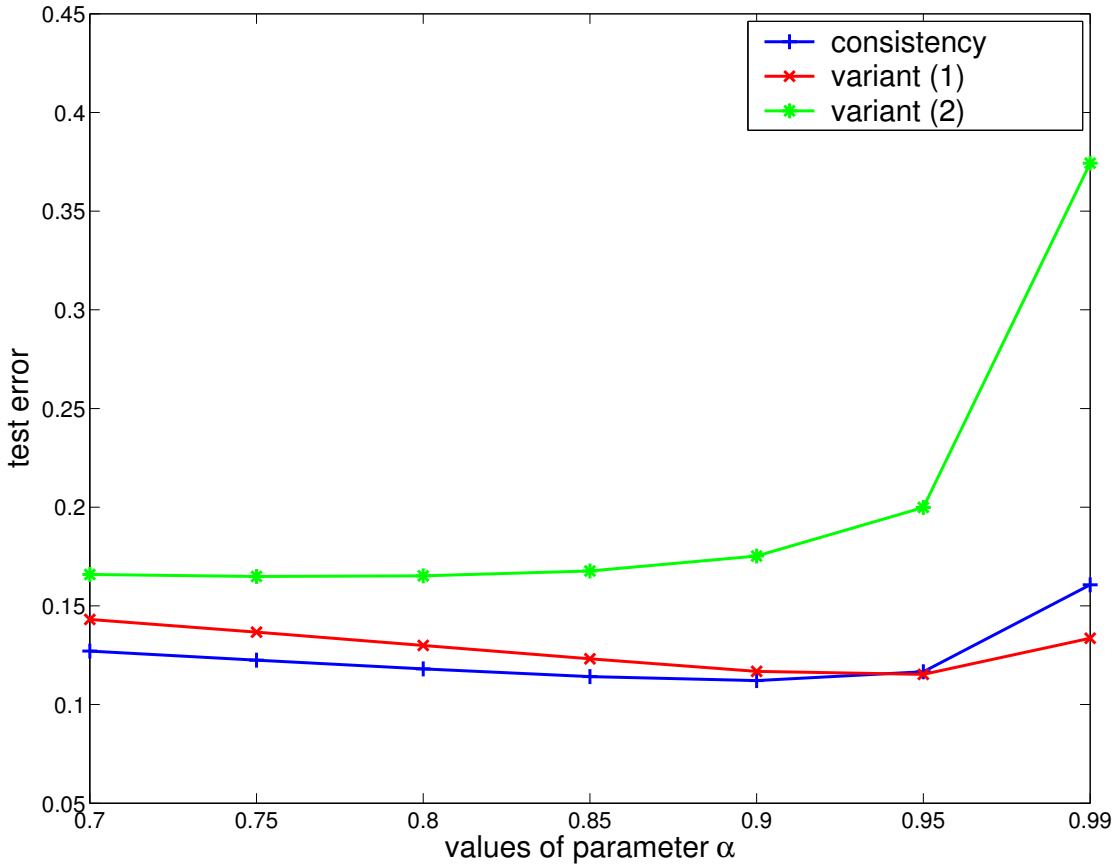


Handwritten Digit Recognition (USPS)



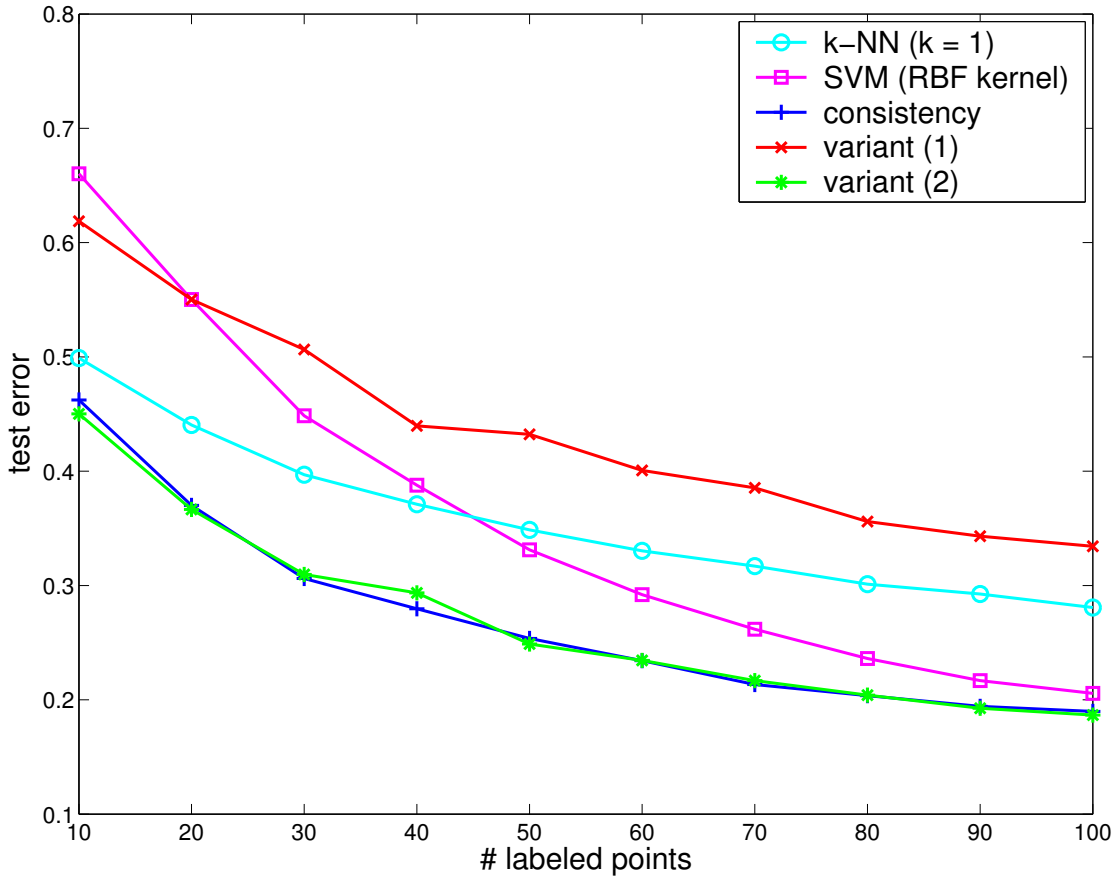
Dimension: 16x16. Size: 9298. ($\alpha = 0.95$)

Handwritten Digit Recognition (USPS)



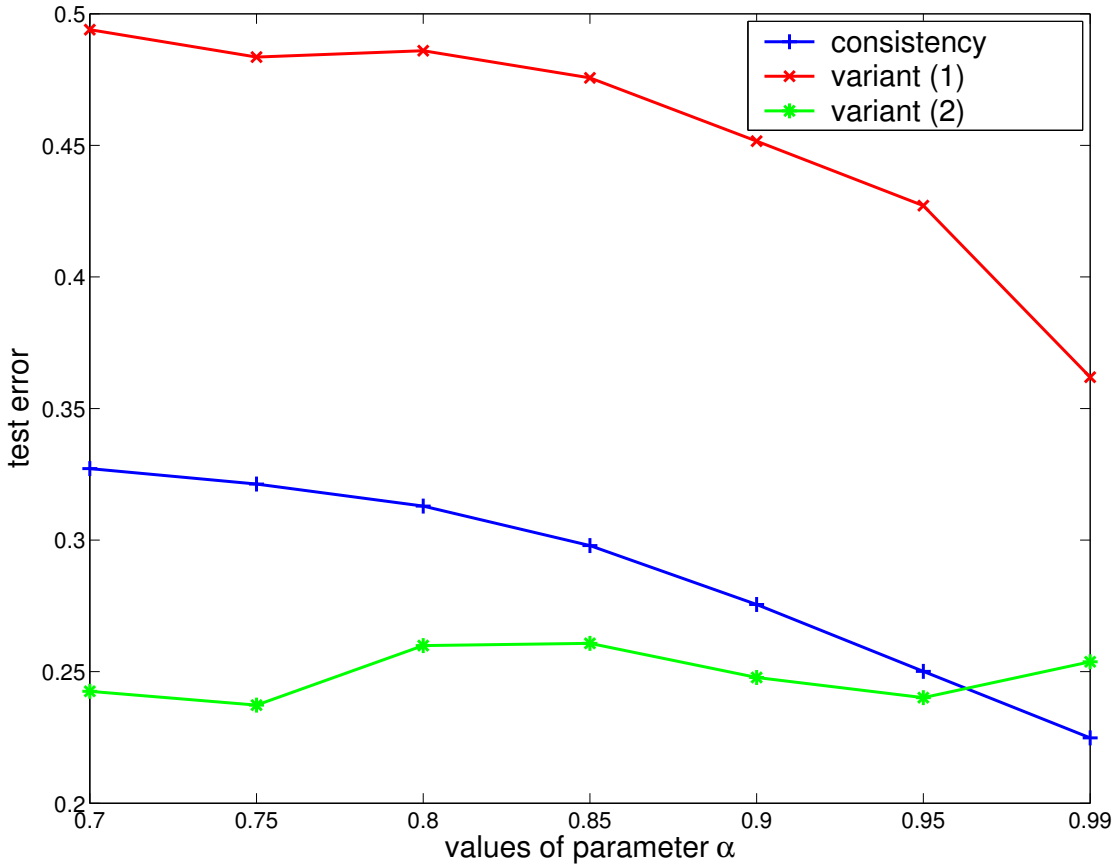
Size of labeled data: $l = 50$.

Text Classification (20-newsgroups)



Dimension: 8014. Size: 3970. ($\alpha = 0.95$)

Text Classification (20-newsgroups)



Size of labeled data: $l = 50$.