Random Projections, Margins, Kernels and Feature Selection

Adithya Pediredla

Rice University
Electrical and Computer Engineering

SVM: Revision

- $f(x_i) = w^T x_i + b$

SVM: Revision

- $f(x_i) = w^T x_i + b$
- Primal: $\min_{w \in \mathcal{R}^d} \|w\|^2 + C \sum_{i}^{N} \max(0, 1 - y_i f(x_i));$

SVM: Revision

- $f(x_i) = w^T x_i + b$
- Primal: $\min_{w \in \mathcal{R}^d} \|w\|^2 + C \sum_i^N \max(0, 1 - y_i f(x_i));$
- Dual: $\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{j,k} \alpha_i \alpha_j y_j y_k (x_j^T x_k);$

  S.T. $0 \leq \alpha_i \leq C; \sum_i \alpha_i y_i = 0, \ \forall i$

SVM: Revision

- $f(x_i) = w^T x_i + b$
- Primal: $\min_{w \in \mathcal{R}^d} \|w\|^2 + C \sum_i^N \max(0, 1 - y_i f(x_i));$
- Dual: $\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{j,k} \alpha_i \alpha_j y_j y_k (x_j^T x_k);$

  S.T. $0 \leq \alpha_i \leq C; \sum_i \alpha_i y_i = 0, \ \forall i$

  only inner products matter

SVM: Revision

- $f(x_i) = w^T x_i + b$
- Primal: $\min_{w \in \mathcal{R}^d} \|w\|^2 + C \sum_i^N \max(0, 1 - y_i f(x_i))$; $\mathcal{O}(nd^2 + d^3)$
- Dual: $\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{j,k} \alpha_i \alpha_j y_j y_k (x_j^T x_k)$; $\mathcal{O}(dn^2 + n^3)$

  S.T. $0 \leq \alpha_i \leq C; \sum_i \alpha_i y_i = 0, \ \forall i$

  only inner products matter

Decreasing computations

- Only inner products matter.

Decreasing computations

- Only inner products matter.
- Can we approximate x_i with z_i so that $\dim(z_i) << \dim(x_i)$ and $x_i^T x_j \approx z_i^T z_j$.

Decreasing computations

- Only inner products matter.
- Can we approximate x_i with z_i so that $\dim(z_i) << \dim(x_i)$ and $x_i^T x_j \approx z_i^T z_j$.
- One way $z_i = Ax_i$.

Decreasing computations

- Only inner products matter.
- Can we approximate x_i with z_i so that $\dim(z_i) << \dim(x_i)$ and $x_i^T x_j \approx z_i^T z_j$.
- One way $z_i = Ax_i$.
  Any comment on rows vs columns of A.

Decreasing computations

- Only inner products matter.
- Can we approximate x_i with z_i so that $\dim(z_i) << \dim(x_i)$ and $x_i^T x_j \approx z_i^T z_j$.
- One way $z_i = Ax_i$.
  Any comment on rows vs columns of A.
- Turns out a random A is good !!

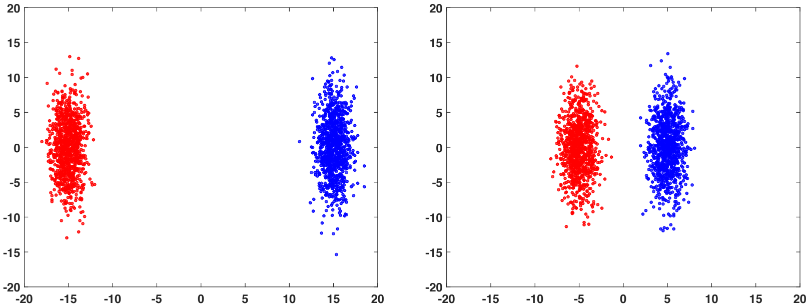
Johnson-Linderstrauss Lemma

- If $d_{new} = \omega(\frac{1}{\gamma^2} \log n)$, relative angles are preserved up to $1 \pm \gamma$.

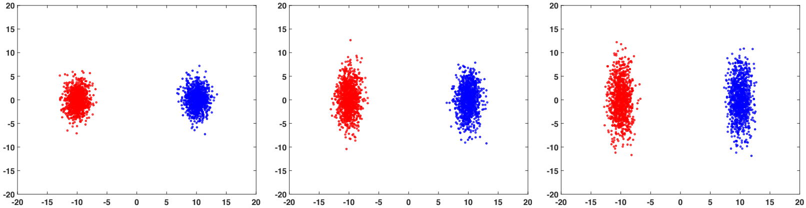
Johnson-Linderstrauss Lemma

- If $d_{new} = \omega(\frac{1}{\gamma^2} \log n)$, relative angles are preserved up to $1 \pm \gamma$.
- How big can γ be?

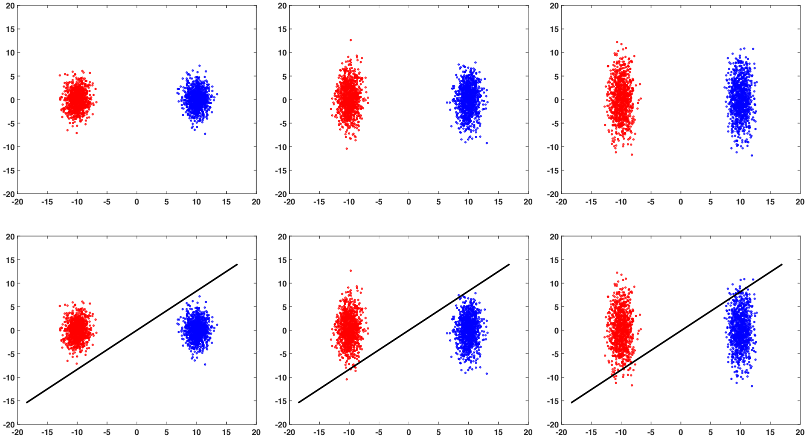
which data set can have higher γ



which data set can have higher γ



which data set can have higher γ



How else can big margin help

A simple weak learner whose speed is proportional to margin.

step 1: Pick random h.

step 2: Evaluate error in step 1.
If error $< \frac{1}{2} - \frac{\gamma}{4}$, stop
else, goto step 1.

How else can big margin help

A simple weak learner whose speed is proportional to margin.

step 1: Pick random h.

step 2: Evaluate error in step 1.

If error $< \frac{1}{2} - \frac{\gamma}{4}$, stop

else, goto step 1.

Bigger the margin, lesser the iterations

Dimensionality reduction: random projection

Coming back to random projection. $A_{d \times D}$

1. Choose columns to be D random orthogonal unit-length vectors.

Dimensionality reduction: random projection

Coming back to random projection. $A_{d \times D}$

1. Choose columns to be D random orthogonal unit-length vectors.
2. Choose each entry in A independently from a standard Gaussian.

Dimensionality reduction: random projection

Coming back to random projection. $A_{d \times D}$

1. Choose columns to be D random orthogonal unit-length vectors.
2. Choose each entry in A independently from a standard Gaussian.
3. Choose each entry in A to be 1 or -1 independently at random.

Dimensionality reduction: random projection

Coming back to random projection. $A_{d \times D}$

1. Choose columns to be D random orthogonal unit-length vectors.
2. Choose each entry in A independently from a standard Gaussian.
3. Choose each entry in A to be 1 or -1 independently at random.

For (2) and (3):

$$Pr_A[(1-\gamma)\|u-v\|^2 \leq \|u'-v'\|^2 \leq (1+\gamma)\|u-v\|^2] \geq 1-2e^{-(\gamma^2-\gamma^3)\frac{d}{4}}$$

Dimensionality reduction: random projection

Coming back to random projection. $A_{d \times D}$

1. Choose columns to be D random orthogonal unit-length vectors.
2. Choose each entry in A independently from a standard Gaussian.
3. Choose each entry in A to be 1 or -1 independently at random.

For (2) and (3):
$Pr_A[(1-\gamma)\|u-v\|^2 \leq \|u'-v'\|^2 \leq (1+\gamma)\|u-v\|^2] \geq 1 - 2e^{-(\gamma^2-\gamma^3)\frac{d}{4}}$

Can we do better?

Can we do better

If $Pr(\text{error} < \epsilon) < \delta$

Can we do better

If $Pr(\text{error} < \epsilon) < \delta$

$d = \mathcal{O}(\frac{1}{\gamma^2}\log(\frac{1}{\epsilon\delta}))$ is sufficient.

Kernel functions

- What if we know that $K(x_1, x_2) = \phi(x_1)\phi(x_2)$?

Kernel functions

- What if we know that $K(x_1, x_2) = \phi(x_1)\phi(x_2)$?
- What if we do not?

Kernel functions

- What if we know that $K(x_1, x_2) = \phi(x_1)\phi(x_2)$?
- What if we do not? Finding Inner products approximately is enough

Kernel functions

- What if we know that $K(x_1, x_2) = \phi(x_1)\phi(x_2)$?
- What if we do not? Finding Inner products approximately is enough
- We need to know the distribution of data set

Mapping-1

Lemma: Consider any distribution over labelled data.

Mapping-1

Lemma: Consider any distribution over labelled data.
Assume $\exists\, w \ni P[\|w \cdot x\| > \gamma] = 0$.

Mapping-1

Lemma: Consider any distribution over labelled data.
Assume $\exists\, w \ni P[\|w \cdot x\| > \gamma] = 0$.
If we draw $z_1, z_2, \ldots z_d$ iid with $d \geq \frac{8}{\epsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]$ then with probability $\geq 1 - \delta$, $\exists\, w' = \text{span}(z_1, z_2, \ldots, z_d) \ni P[\|w' \cdot x\| > \gamma/2] < \epsilon$

Mapping-1

Lemma: Consider any distribution over labelled data.
Assume $\exists\, w \ni P[\|w \cdot x\| > \gamma] = 0$.
If we draw $z_1, z_2, \ldots z_d$ iid with $d \geq \frac{8}{\epsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]$ then with probability $\geq 1 - \delta$, $\exists\, w' = \text{span}(z_1, z_2, \ldots, z_d) \ni P[\|w' \cdot x\| > \gamma/2] < \epsilon$

Therefore, if $\exists w$ in $\phi-$space, by sampling $x_1, x_2, \ldots x_n$, we are guaranteed:
$w' = \alpha_1\phi(x_1) + \alpha_2\phi(x_2) + \cdots + \alpha_d\phi(x_d)$
Hence,
$w'\phi(x) = \alpha_1 K(x, x_1) + \alpha_2 K(x, x_2) + \ldots \alpha_d K(x, x_d)$;

Mapping-1

Lemma: Consider any distribution over labelled data.
Assume $\exists\, w \ni P[\|w \cdot x\| > \gamma] = 0$.
If we draw $z_1, z_2, \ldots z_d$ iid with $d \geq \frac{8}{\epsilon}\left[\frac{1}{\gamma^2} + \ln \frac{1}{\delta}\right]$ then with probability $\geq 1 - \delta$, $\exists\, w' = \text{span}(z_1, z_2, \ldots, z_d) \ni P[\|w' \cdot x\| > \gamma/2] < \epsilon$

Therefore, if $\exists w$ in $\phi-$space, by sampling $x_1, x_2, \ldots x_n$, we are guaranteed:
$w' = \alpha_1 \phi(x_1) + \alpha_2 \phi(x_2) + \cdots + \alpha_d \phi(x_d)$
Hence,
$w' \phi(x) = \alpha_1 K(x, x_1) + \alpha_2 K(x, x_2) + \ldots \alpha_d K(x, x_d)$;

If we define $F_1(x) = (K(x, x_1), \ldots, K(x, x_d))$; then with high probability the vector $(\alpha_1, \ldots \alpha_d)$ is an approximate linear separator.

Mapping-2

- We can normalize $K(x, x_i)$ and get better bounds.

Mapping-2

- We can normalize $K(x, x_i)$ and get better bounds.
- Compute $K = U^T U$;

Mapping-2

- We can normalize $K(x, x_i)$ and get better bounds.
- Compute $K = U^T U$;
- Compute $F_2(x) = F_1(x)U^{-1}$.

Mapping-2

- We can normalize $K(x, x_i)$ and get better bounds.
- Compute $K = U^T U$;
- Compute $F_2(x) = F_1(x)U^{-1}$.
- F_2 is linearly separable with error at most ϵ at margin $\gamma/2$

Key take aways

- Inner products are enough.
- Random projections are good.
- Higher the margin, lower the dimension.
- If okay with error, we can project to much lower dimension.
- While using Kernels, randomly drawn data points act as good features.