To Winnow

Winnowing Subspaces
How I do it
A formal definition

To *winnow*:

- to remove (chaff from grain) by a current of air
- to get rid of (something undesirable or unwanted)

From Anglo-Saxon “windwian”
The **original Winnow algorithm**

- Online alg. for learning disjunctions
- Mistake bound **logarithmic** in number of features winnows a large number of features
- Will morph into algorithm for **winnowing subspaces**
Disjunctions as linear threshold functions

- 2 out of 5 literal monotone disjunction \( v_1 \lor v_3 \)
- Represented as \( d = (1, 0, 1, 0, 0)^\top \)
- Label for instance \( x = (0, 1, 1, 0, 0)^\top \)
  \[
  \begin{cases} 
  +1 & \text{if } d \cdot x \geq \frac{1}{2} \\
  -1 & \text{otherwise} 
  \end{cases}
  \]
- Alg. receives sequence of examples online
  \[
  (x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)
  \]
  \[
  \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T
  \]

instances \([0, 1]^n\), labels and predictions are \(\pm 1\)
Original Winnow algorithm

Initialize $\mathbf{w}_1 = \mathbf{w}_0 (1, 1, \ldots, 1)\top$

for $t = 1$ to $T$ do

Receive instance $\mathbf{x}_t \in [0, 1]^n$

Predict with

$$
\hat{y}_t = \begin{cases} 
+1 & \text{if } \mathbf{w}_t \cdot \mathbf{x}_t \geq \theta \\
-1 & \text{otherwise}
\end{cases}
$$

Receive label $y_t \in \{+1, -1\}$

Update

$$
\mathbf{w}_{t+1,i} = \begin{cases} 
\mathbf{w}_{t,i} & \text{if no mistake} \\
\mathbf{w}_{t,i} e^{\eta y_t x_t,i} & \text{if mistake}
\end{cases}
$$

end for

Like perceptron alg., except multiplicative update
**Thm** If examples consistent with $k$ out of $n$ literal monotone disjunction, then properly tuned Winnow make at most $O(k \ln n)$ mistakes.

- Mistake bound logarithmic in dimension $n$.
- Perceptron alg. can make $\Omega(nk)$ mistakes.
Morphing into the matrix case?

- What variables?
- What dot product?
- What corresponds to disjunctions?
- What happens to the exponential form of weights - soft max?
Lifting it to the matrix setting

<table>
<thead>
<tr>
<th></th>
<th>states</th>
<th>mixture states</th>
</tr>
</thead>
</table>
| original | \[
\begin{pmatrix}
0 \\
1 \\
0
\end{pmatrix}
\] | \[
\begin{pmatrix}
0.1 \\
0.3 \\
0.6
\end{pmatrix}
\] |
| new     | \[
\mathbf{u}\mathbf{u}^\top
\] \[
\mathbf{W} = \mathbf{U}\begin{pmatrix}
0.1 & 0 & 0 \\
0 & 0.3 & 0 \\
0 & 0 & 0.6
\end{pmatrix}\mathbf{U}^\top
\] |
Visualization of symmetric matrices

- As ellipses - affine transformations of the unit ball

\[ \text{Ellipse} = \{ \mathbf{Wu} : \| \mathbf{u} \|_2 = 1 \} \]
Ellipses cont.

- Eigenvectors form the axes and eigenvalues their lengths.
Dyads

Degenerate ellipses

- One eigenvalue one
- All others zero
Linear combinations and mixtures of dyads

- **Symmetric matrices** are linear combinations of dyads

  \[ \mathbf{U}^\top = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \]

  - orthogonal mat. of eigenvectors
  - diagonal mat. of real eigenvalues
  - real eigenvalues
  - dyads

- **Positive definite matrices**
  - Eigenvalues are non-negative

- **Density matrices** are mixtures of dyads
  - Eigenvalues form probability vector
Mixtures of dyads

- Many mixtures lead to same density matrix

\[ 0.2 + 0.3 + 0.5 = \begin{pmatrix} 0.35 & 0.15 \\ 0.15 & 0.65 \end{pmatrix} = \begin{pmatrix} 0.29 \\ 0.71 \end{pmatrix} \]

- There always exists a decomposition into \( n \) dyads that correspond to eigenvectors
- Uncertainty about dyad expressed as density matrix
- We have a Bayes rule for density matrices
View the symmetric positive definite matrix $W$ as a covariance matrix of some random cost vector $c \in \mathbb{R}^n$

$$W = \mathbb{E} \left( (c - \mathbb{E}(c))(c - \mathbb{E}(c))^\top \right)$$

The variance along any vector $u$ is

$$V(c^\top u) = \mathbb{E} \left( (c^\top u - \mathbb{E}(c^\top u))^2 \right)$$

$$= u^\top \mathbb{E} \left( (c - \mathbb{E}(c))(c - \mathbb{E}(c))^\top \right) u$$

Variance as trace

$$u^\top Wu = \text{tr}(u^\top Wu) = \text{tr}(W uu^\top) \geq 0$$
Curve of the ellipse is plot of vector $Wu$, where $u$ is unit vector.
The outer figure eight is direction $u$ times the variance $u^T Wu$.
For an eigenvector, this variance equals the eigenvalue and touches the ellipse.
3 dimensional variance plots
What dot product?

\[
\text{tr}(W X) = \text{tr}(\sum_i \omega_i w_i w_i^\top X)
\]
\[
= \sum_i \omega_i \text{tr}(w_i w_i^\top X)
\]
\[
= \sum_i \omega_i w_i^\top X w_i
\]

Measurement in quantum physics

- Dyad \( uu^\top \) is state
- Density matrix \( W \) is mixture state
- Instance matrix \( X \) is instrument
- \( \text{tr}(W X) \) is expected outcome

variance along eigendirs
So far

<table>
<thead>
<tr>
<th></th>
<th>states</th>
<th>mixture states</th>
<th>dot product</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>$\begin{pmatrix} 0 \ 1 \ 0 \end{pmatrix}$</td>
<td>$w = \begin{pmatrix} 0.1 \ 0.3 \ 0.6 \end{pmatrix}$</td>
<td>$w \cdot x$</td>
</tr>
<tr>
<td>new</td>
<td>$uu^\top$</td>
<td>$W = U \begin{pmatrix} 0.1 &amp; 0 &amp; 0 \ 0 &amp; 0.3 &amp; 0 \ 0 &amp; 0 &amp; 0.6 \end{pmatrix} U^\top$</td>
<td>$\text{tr}(WX)$</td>
</tr>
</tbody>
</table>
What corresponds to disjunctions

- **Disjunctions**

\[(1, 0, 1, 0, 0)^\top \cdot (x_1, x_2, x_3, x_4, x_5)^\top = x_1 + x_3\]

**Sum k components of x**

- **Projections matrices** \( P = \sum_{i=1}^{k} u_i u_i^\top \)

\[\text{tr}(PX) = \sum_{i=1}^{k} u_i^\top X u_i\]

**Sum variance along k directions**

\[\left( X_1, y_1 \right), \left( X_2, y_2 \right), \ldots, \left( X_T, y_T \right)\]

\(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\)

Label \( y_t \) is +1 if trace is at least \( \frac{1}{2} \) and -1 otherwise
Thresholding the trace for $\mathbf{xx}^\top$ instances
Symmetric Matrix Winnow

Initialize $W_1 = w_0 \cdot I_{n \times n}$

for $t = 1$ to $T$ do

Receive instance $X_t \in \mathbb{R}^{n \times n}$ with eigenvalues in $[0, 1]$  

Predict with 

$$\hat{y}_t = \begin{cases} 
+1 & \text{if } \text{tr}(W_t X_t) \geq \theta \\
-1 & \text{otherwise}
\end{cases}$$

Receive label $y_t$

Update 

$$W_{t+1} = \begin{cases} 
W_t & \text{if no mistake} \\
\exp(\log W_t + \eta y_t X_t) & \text{if mistake}
\end{cases}$$

end for

$\exp$ and $\log$ are spectral functions

In normalized version, trace normalized to one
What else?

- **Same** $O(k \ln n)$ mistake bound if examples consistent with $k$-dimensional subspace
  - Dubbed *free matrix lunch*

- Generalizes to arbitrary matrices
  - use mixtures of $uv^\top$ and SVD

- Key tool in analysis is **quantum relative entropy**

\[ \Delta(W, V) = \text{tr}(W(\log W - \log V)) \]