Supervised learning: Given a few examples find a predictor function that works for new examples.

We are given a training set. We’ll primarily talk about batch / offline supervised classification where training data is given upfront.

Training set: \( S_n = \{(x^i, y^i), i = 1, ..., n\}, (x^i, y^i) \sim p^* \)

We select \( \hat{h}: X = R^d \to Y = \{-1, 1\} \)

Goal is to minimize generalization error (risk):

\[
R(\hat{h}) = E(x, y) \sim p^* \left\{ Loss(y, \hat{h}(x)) \right\}
\]

(expected value of loss on pairs sampled over \( p^* \))

\[
Loss(y, \hat{h}(x)) = \begin{cases} 
1, & \text{if } y \neq y' \\
0, & \text{o. w.}
\end{cases}
\]

Generative approach to solve the supervised learning problem

If I knew \( p^* \) I could perform optimal. I could evaluate what the generalization error is and find an \( \hat{h} \) that would minimize it.

Estimate \( \hat{p}(x, y) \) based on the training set \( S_n \). We need constraints for estimating.

Once we have a guess on \( p^* \) we’ll use a predictor \( \hat{h}(\cdot) = \arg\min_h E(x, y) \sim \hat{p} \left\{ Loss(y, h(x)) \right\} \)

Discriminative approach

\[
\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} Loss\left(y^i, h(x^i)\right)
\]

\( h \in H = \text{set of classifiers} \)

We want to find an \( h \) minimize \( \hat{R}_n(h) \)

Use \( \hat{h}(\cdot) = \arg\min_h \hat{R}_n(h) \)
Linear classifiers

\[ X = \mathbb{R}^d \]
\[ Y = \{-1, 1\} \]
\[
h(x; \theta, \theta_0) = \text{sign}(\theta x + \theta_0) = \begin{cases} +1, & \text{if } \theta x + \theta_0 > 0 \\ -1, & \text{otherwise} \end{cases}
\]

You can get arbitrarily complex classifiers if you know how to handle linear classifiers.

\[ \theta \tilde{x} + \theta_0 = 0 \iff \theta(\tilde{x} - \tilde{x}_0) = 0 \implies \theta \tilde{x}_0 = -\theta_0 \implies \theta_0 = -\theta \tilde{x}_0 \]

Distance to boundary \( \frac{y(\theta x)}{||\theta||} \), where \( ||\theta|| = \text{norm of theta} \)

Figure 1: Geometry of linear discriminant functions (Bishop Figure 4.1).

TODO: Put graph in

Training without errors

Assumption 1: Training examples are linearly separable with margin \( \gamma \):

\[ \exists \theta^* \text{ s.t. } \forall i, \frac{y^i \theta^* x^i}{||\theta||} > \gamma, \gamma > 0, i = 1, ..., n \]

This says that all training examples are at least distance \( \gamma \) from the boundary.

Put another way, this means the examples are linearly separable.

Assumption 2: Training examples are bounded by a sphere/circle of radius \( r \): \( ||x^{(i)}|| \leq r, i = 1, ..., n \)

Perceptron algorithm

- Start at step 0: \( \theta^{(0)} = 0 \) (vector)
- Cycle through training samples correcting errors
Thus, if our two assumptions hold, then our perception algorithm makes at most $\frac{r^2}{\gamma^2}$ mistakes.

The number of mistakes does not depend on the number of training examples or on the dimension of $X$, $\dim(X)$.

**Training without errors online**

Assumption 1: Training examples are linearly separable with margin $\gamma$:

$$\exists \theta^* \text{ s.t. } \forall i \gamma i \theta^* x^i > \gamma, \gamma > 0, i = 1, ..., \infty$$

Assumption 2: Training examples are bounded by a sphere/circle or radius $r$: $\|x^i\| \leq r$, $i = 1, ..., \infty$

**Perceptron algorithm online**

- Start at step 0: $\theta^{(0)} = 0$ (vector)
- Cycle through training samples correcting errors
- If $y^{(i)}(\theta^{(k)} x^{(i)}) \leq 0$ (mistake), then $\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$

Once again, if our two assumptions hold, then our perception algorithm makes at most $\frac{r^2}{\gamma^2}$ mistakes.

Why the $\frac{r^2}{\gamma^2}$ bound?

$$\cos(\theta^k, \theta^*) = \frac{\theta^k \theta^*}{\|\theta^k\| \times \|\theta^*\|} \geq k \gamma$$

where $\theta^k$ is theta after $k$ updates and theta star is the theta we assume exists

**Step 1:** Show that as we keep updating $\frac{\theta^k \theta^*}{\|\theta^k\|} \geq k \gamma$

$$\frac{\theta^k \theta^*}{\|\theta^*\|} = \frac{\theta^{k-1} + y^i x^i \times \theta^*}{\|\theta^*\|} = \frac{\theta^{k-1} \times \theta^*}{\|\theta^*\|} + \frac{y^i x^i \times \theta^*}{\|\theta^*\|} \geq \frac{y^i x^i \times \theta^*}{\|\theta^*\|} \geq \gamma$$

$$\Rightarrow \frac{\theta^k \theta^*}{\|\theta^*\|} \geq k \gamma$$

**Step 2:** Norm of our parameter vector does not increase too high: $\|\theta^k\|^2 \leq k r^2$

- We only update based on a mistake. We are correcting mistakes, which keeps the norm in check.

$$\cos(\theta^k, \theta^*) = \frac{\theta^k \theta^*}{\|\theta^k\| \times \|\theta^k\|} \geq \frac{k \gamma}{\sqrt{k} \times r} = \sqrt{k} \frac{\gamma}{r}$$

Why use the perception algorithm when we can find the maximum margin linear separator directly? We know $\gamma$. 
Maximum margin linear separator

\[
\begin{align*}
\text{minimizing} & \quad \frac{1}{2} \| \theta \|^2 \\
y^i \theta x^i & \geq 1, \forall i
\end{align*}
\]

Unique answer