Supervised Learning :: CHEAT SHEET

Linear hard-margin SVM

For labeled data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$, with $y^{(i)} \in \{-1, +1\}$:

- Assume linear separation by $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$, such that all $+$-observations are in the positive halfspace $\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) > 0\}$ and all $-$-observations are in the negative halfspace $\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) < 0\}$.
- For a linear separating hyperplane, we have

$$y^{(i)} \underbrace{\left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \theta_0\right)}_{=f\left(\mathbf{x}^{(i)}\right)} > 0 \quad \forall i \in \{1, 2, ..., n\}.$$

- $$d\left(f, \mathbf{x}^{(i)}\right) = \frac{y^{(i)} f\left(\mathbf{x}^{(i)}\right)}{\|\boldsymbol{\theta}\|} = y^{(i)} \frac{\boldsymbol{\theta}^T \mathbf{x}^{(i)} + \theta_0}{\|\boldsymbol{\theta}\|}$$

computes the (signed) distance to the separating hyperplane $f(\mathbf{x}) = 0$, positive for correct classifications, negative for incorrect.

- The distance of f to the whole dataset $\mathcal{D}$ is the smallest distance $\gamma = \min_i \left\{ d\left(f, \mathbf{x}^{(i)}\right) \right\}$, which represents the **safety margin**. It is positive if f separates and we want to maximize it.

$$\begin{aligned} \max_{\boldsymbol{\theta}, \theta_0} \ & \gamma \\ \text{s.t.} \ & d\left(f, \mathbf{x}^{(i)}\right) \geq \gamma \quad \forall i \in \{1, \ldots, n\}. \end{aligned}$$

Primal linear hard-margin SVM:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ \text{s.t.} \quad & y^{(i)}\left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \theta_0\right) \geq 1 \quad \forall i \in \{1, \ldots, n\} \end{aligned}$$

This is a convex quadratic program.

Support vectors: All instances $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$ with minimal margin $y^{(i)} f\left(\mathbf{x}^{(i)}\right) = 1$, fulfilling the inequality constraints with equality. All have distance of $\gamma = 1/\|\boldsymbol{\theta}\|$ from the separating hyperplane.

The Lagrange function of the SVM optimization problem is

$$\begin{aligned} L(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{i=1}^n \alpha_i \left[y^{(i)} \left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \theta_0 \right) - 1 \right] \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i \in \{1, \ldots, n\}. \end{aligned}$$

The **dual** form of this problem is $\max_{\alpha} \min_{\boldsymbol{\theta}, \theta_0} L(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha})$.

We find the stationary point of $L(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\theta}, \theta_0$ and obtain

$$\boldsymbol{\theta} = \sum_{i=1}^n \alpha_i y^{(i)} \mathbf{x}^{(i)}, 0 = \sum_{i=1}^n \alpha_i y^{(i)} \quad \forall i \in \{1, \ldots, n\}.$$

Dual linear hard-margin SVM:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \ & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} \left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle \\ \text{s.t.} \ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \\ & \alpha_i \geq 0 \ \forall i \in \{1, \ldots, n\}, \end{aligned}$$

In matrix notation with $\boldsymbol{K} := \mathbf{X}\mathbf{X}^T$:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \ & \mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \operatorname{diag}(\mathbf{y}) \boldsymbol{K} \operatorname{diag}(\mathbf{y}) \boldsymbol{\alpha} \\ \text{s.t.} \ & \boldsymbol{\alpha}^T \mathbf{y} = 0, \\ & \boldsymbol{\alpha} \geq 0, \end{aligned}$$

Solution (if existing):

$$\hat{\theta} = \sum_{i=1}^n \hat{\alpha}_i y^{(i)} \mathbf{x}^{(i)}, \quad \theta_0 = y^{(i)} - \left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle.$$

Linear Soft-Margin SVM

Allow violations of the margin constraints via slack vars $\zeta^{(i)} \geq 0$

$$y^{(i)} \left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \boldsymbol{\theta}_0 \right) \geq 1 - \zeta^{(i)}$$

Now we have two distinct and contradictory goals:

- Maximize the margin.
- Minimize margin violations.

Primal linear soft-margin SVM:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\theta}_0, \zeta^{(i)}} \ & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \zeta^{(i)} \\ \text{s.t.} \ & y^{(i)} \left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \boldsymbol{\theta}_0 \right) \geq 1 - \zeta^{(i)} \quad \forall i \in \{1, \ldots, n\}, \\ \text{and} \ & \zeta^{(i)} \geq 0 \quad \forall i \in \{1, \ldots, n\}, \end{aligned}$$

where the constant $C > 0$ controls trade-off between the two conflicting objectives of maximizing the size of the margin and minimizing the frequency and size of margin violations.

Dual linear soft-margin SVM:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \ & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} \left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle \\ \text{s.t.} \ & 0 \leq \alpha_i \leq C, \forall i \in \{1, \ldots, n\} \quad \text{and} \quad \sum_{i=1}^n \alpha_i y^{(i)} = 0 \end{aligned}$$

- Non-SVs have $\alpha_i = 0$ ($\Rightarrow \mu_i = C \Rightarrow \zeta^{(i)} = 0$) and can be removed from the problem without changing the solution. Their margin $yf(\mathbf{x}) \geq 1$. They are always classified correctly and are never inside of the margin.
- SVs with $0 < \alpha_i < C$ ($\Rightarrow \mu_i > 0 \Rightarrow \zeta^{(i)} = 0$) are located exactly on the margin and have $yf(\mathbf{x}) = 1$.
- SVs with $\alpha_i = C$ have an associated slack $\zeta^{(i)} \geq 0$. They can be on the margin or can be margin violators with $yf(\mathbf{x}) < 1$ (they can even be misclassified if $\zeta^{(i)} \geq 1$).

Regularized ERM representation with hinge loss:

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right); \ L(y, f(\mathbf{x})) = \max(1 - yf(\mathbf{x}), 0)$$

Optimization

Algorithm 1 Stochastic subgradient descent (without intercept θ_0)

```
for t = 1, 2, ... do
    Pick step size α
    Randomly pick an index i
    If y^(i) f(x^(i)) < 1 set θ^[t+1] = (1 − λα)θ^[t] + α y^(i) x^(i)
    If y^(i) f(x^(i)) ≥ 1 set θ^[t+1] = (1 − λα)θ^[t]
end for
```

Algorithm 2 Pairwise coordinate ascent in the dual

```
Initialize α = 0 (or more cleverly)
for t = 1, 2, ... do
    Select some pair α_i, α_j to update next
    Optimize dual w.r.t. α_i, α_j, while holding α_k (k ≠ i, j) fixed
end for
```

Supervised Learning :: CHEAT SHEET

Kernel

Kernel = Feature Map + Inner product

Mercer Kernel

A **(Mercer) kernel** on a space $\mathcal{X}$ is a continuous function

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

of two arguments with the properties

- Symmetry: $k(\mathbf{x}, \tilde{\mathbf{x}}) = k(\tilde{\mathbf{x}}, \mathbf{x})$ for all $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$.
- Positive definiteness: For each finite subset $\left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$ the **kernel Gram matrix** $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is positive semi-definite.

Reproducing property: for all kernels, there must exist a Hilbert space, where a map ϕ of this space satisfies $k(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}) \rangle$. The space is called **reproducing kernel Hilbert space** (RKHS).

Typical Kernels

A kernel can be constructed from other kernels k_1 and k_2:

- For $\lambda \geq 0$, $\lambda \cdot k_1$ is a kernel.
- $k_1 + k_2$ is a kernel.
- $k_1 \cdot k_2$ is a kernel (thus also k_1^n).

Useful kernels:

- Every constant function taking a non-negative value.
- **Linear kernel**: $k(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbf{x}^\top \tilde{\mathbf{x}}$.
- **Polynomial kernel**: $k(\mathbf{x}, \tilde{\mathbf{x}}) = (\mathbf{x}^\top \tilde{\mathbf{x}} + b)^d$, for $b \geq 0, d \in \mathbb{N}$.

$$\phi(\mathbf{x}) = \left(\sqrt{\binom{d}{k_1, \dots, k_{p+1}}} x_1^{k_1} \dots x_p^{k_p} b^{k_{p+1}/2} \right)_{k_i \geq 0, \sum_i k_i = d}$$

- **Gaussian kernel**: $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|^2}{2\sigma^2})$ or $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^2)$, $\gamma > 0$

Dual kernelized soft-margin SVM:

$$\max_\alpha \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \forall\, i \in \{1, \dots, n\} \quad \text{and} \quad \sum_{i=1}^n \alpha_i y^{(i)} = 0$$

Kernel representation of separating hyperplane:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y^{(i)} k(\mathbf{x}^{(i)}, \mathbf{x}) + \theta_0$$

Hyperparameters of SVM

SVMs are somewhat sensitive to its hyperparameters and should always be tuned.

- The choice of C, the choice of the kernel, the kernel parameters are all up to the user.
- Small C allows for margin-violating points in favor of a large margin.
- Large C penalizes margin violators, decision boundary is more wiggly.