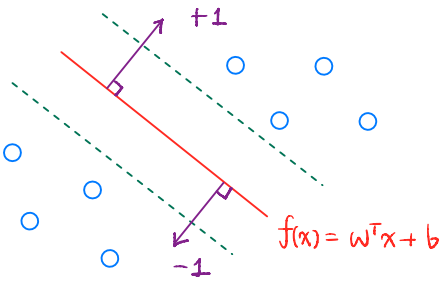
Support Vector Machine (SVM)

단순 logistic regression 보다는 margin을 최대화 하는 decision boundary 찾는 것 (제약사항)

- find a decision boundary with the maximum margin
  - margin: distance from the boundary to the closest training points
- supervised learning model for classification (and regression as SVR)
- linear decision function
  - $f(x) = w^\top x + b$
  - predict by $\text{sign}(f(x))$



Maximum-margin (hard margin)

- separable case
  - constraints: $y_i(w^\top x_i + b) \geq 1$
  - maximize margin $\frac{2}{|w|}$ is equivalent to minimizing $|w|^2$

평면 $w^T x + b = c_1$ 사이의 거리 공식 → margin = $\frac{|c_1 - c_2|}{\|w\|}$

$w^T x + b = c_2$ $c_1 = 1$ $c_2 = -1$

- primal optimization
  - $\min_{w,b}; \frac{1}{2}|w|^2$
  - subject to $y_i(w^\top x_i + b) \geq 1$

$\|w\|^2 \downarrow \rightarrow$ margin $\uparrow$

hard margin: 모든 data가 margin 밖에 있도록

$y \in \{-1, +1\}$

- support vectors
  - the points that lie on the margin boundaries
  - they "support" the optimal hyperplane
  - only these points determine the solution

SVM은 "모든 sample이 margin 밖에 있어야 한다"는 제약이 있는 optimization 문제

$y_i(w^T x_i + b) \geq 1$

Soft margin (non-separable)

현실적으로 hard margin 불가능할 때가 많음 → 약간의 위반량 허용

- allow violations via slack variables ξ_i
  - constraints: $y_i(w^\top x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$
- objective
  - $\min_{w,b}; \frac{1}{2}|w|^2 + C\sum_i \xi_i$

C ↑ → 위반량의 가중치 ↑ → 더 강하게 penalty

C ↓ → 위반량의 가중치 ↓ → 위반량 허용 ↑ w/ large margin

    - e.g., ReLU, Leaky ReLU, GELU
- compensate for ReLU's sparsity
    - about half of activations are zero
    - scaling by 2 → help preserve activation variance & gradient flow

---

Logit & Softmax

Logit

$\text{logit}(p) = \log \frac{p}{1-p}$

- logarithm of the odds of the event
    - raw & unnormalized scores output by the model
    - pre-softmax outputs whose differences correspond to log-odds between classes
- interpreted as evidence for each class

Softmax

$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$

- convert logits into a probability distribution
    - output is positive and sums to 1
    - preserve relative differences between logits
- smooth and differentiable
    - compatible with likelihood-based objectives
- e^z grows extremely fast
    - if some logit z is large, e^z can exceed floating-point range → overflow (`inf`)

Log-sum-exp trick

exp 각 항에서 최댓값 빼 줌 → overflow 방지

$\log \sum_j e^{z_j} = m + \log \sum_j e^{z_j - m}, m = \max_j z_j$ 최댓값 빼줌

- subtract the maximum logit to avoid overflow
  - keep exponentials in a safe range
  - $\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$
  - $z' = z - \max_j z_j$
  - $\text{softmax}(z) = \text{softmax}(z')$

$$\frac{e^{z_i - c}}{\sum_j e^{z_j - c}}$$

Gumbel Softmax

$y_i = \frac{\exp((\log p_i + g_i)/\tau)}{\sum_j \exp((\log p_j + g_j)/\tau)}$

- differentiable approximation to categorical sampling
  - sampling from a categorical distribution is non-differentiable
  - differentiable with respect to logits
- use cases
  - discrete latent variables
  - VQ-VAE style model

Entropy

$H(p) = -\sum_i p_i \log p_i$

- measure uncertainty of a distribution
- defined as the expected amount of surprise
  - surprisal of an event: $-\log p(x)$
- entropy ↑ → uncertainty ↑ → ≈ uniform distibution