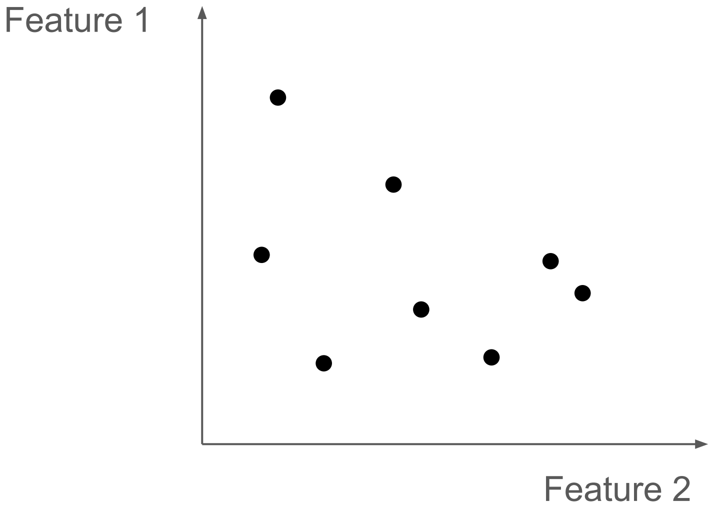
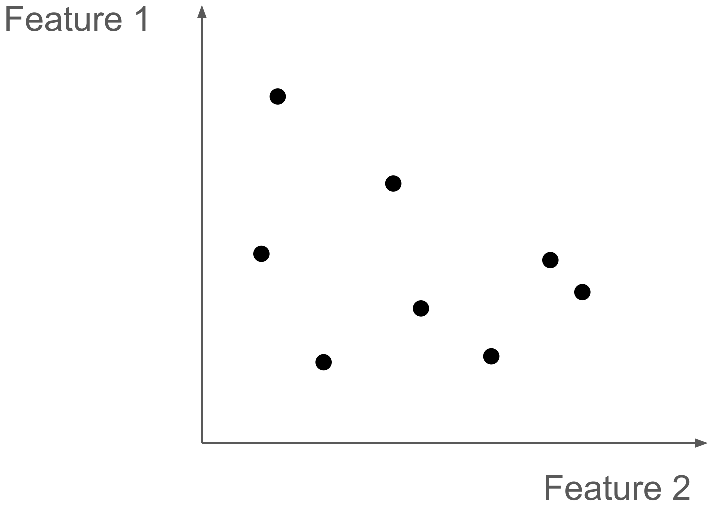
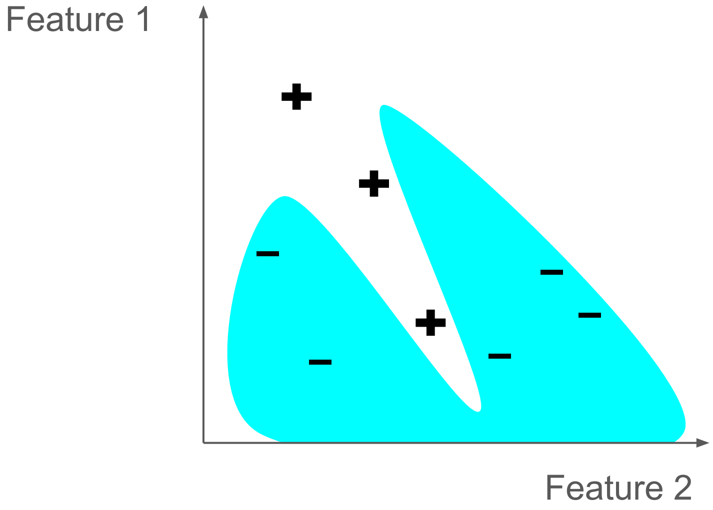

Feature 1
Feature 2

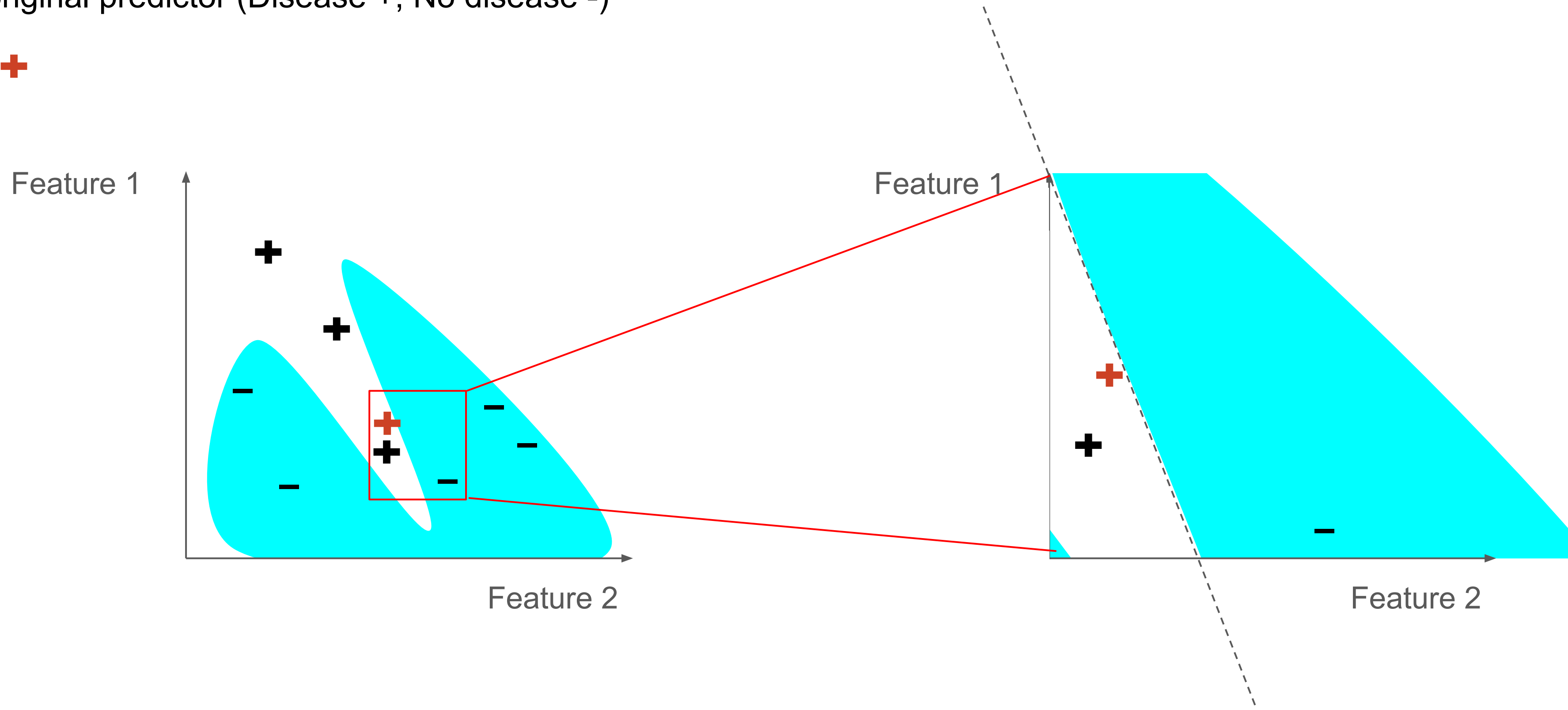


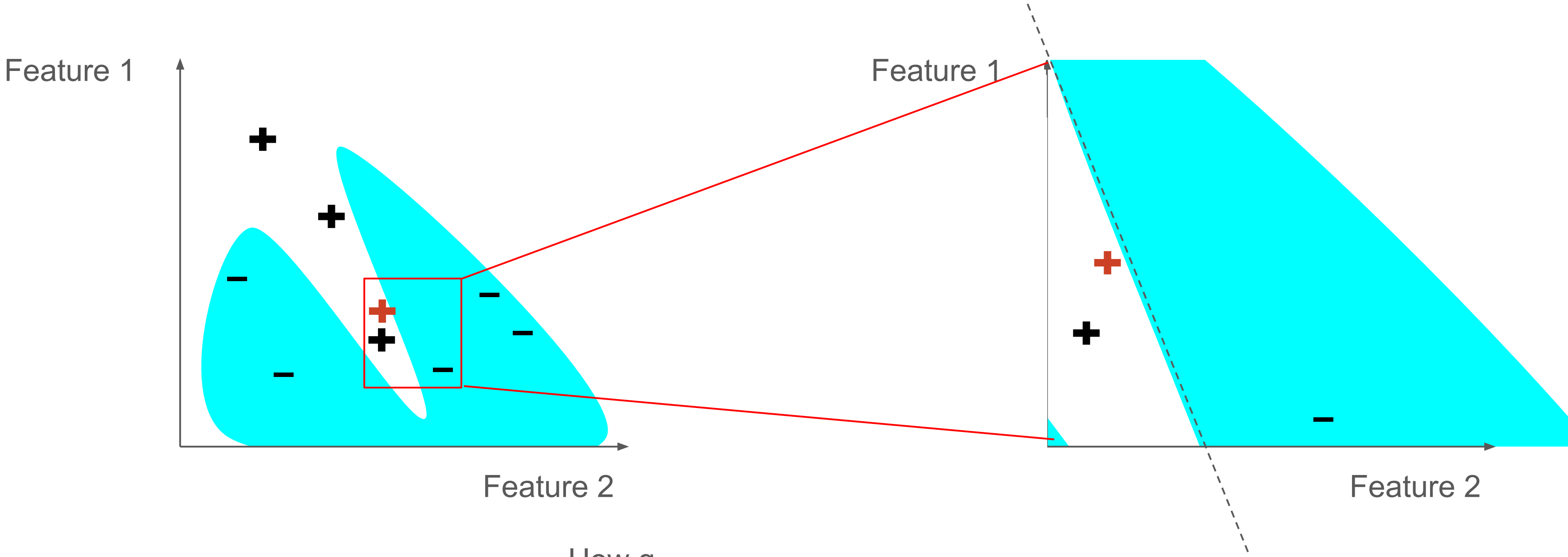
f: an original predictor (Disease +, No disease -)



f: an original predictor (Disease +, No disease -)

f(x) = +





How g approximate to f

Complexity simple model

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

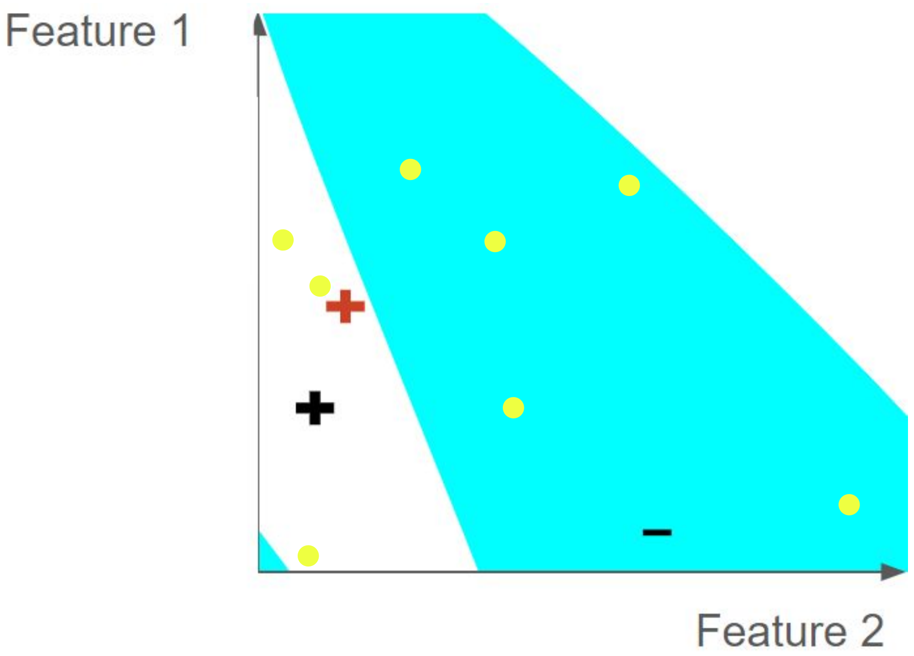
Family Simple models

Complex model

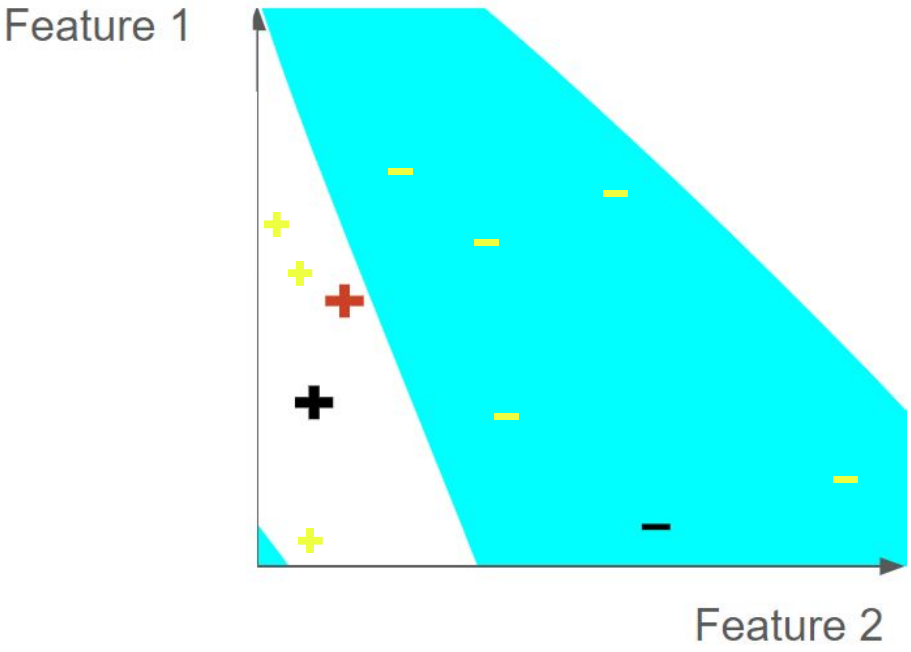
Simple model

Proximity

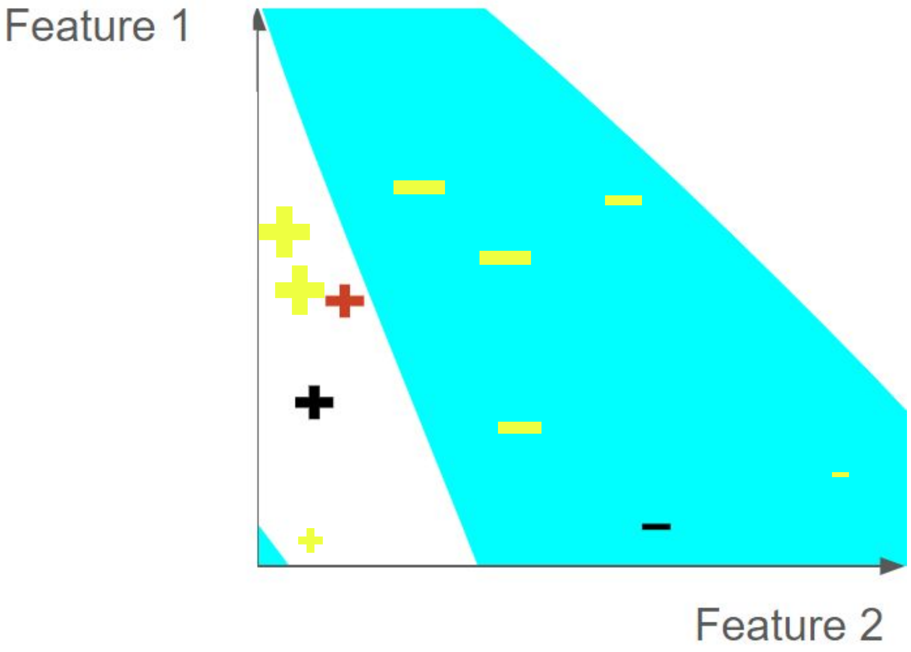
z = ● Perturbation of x



f(z)



$\pi_x(z)$



1. ***x'*** *(interpretable representation)*: This binary vector is a human-understandable version of the actual features used by the original model.
2. **z'** (perturbed sample): a fraction of non zero elements of x'.

model. For example, a possible *interpretable representation* for text classification is a binary vector indicating the presence or absence of a word, even though the classifier may use more complex (and incomprehensible) features such as word embeddings. Likewise for image classification, an *interpretable representation* may be a binary vector indicating the "presence" or "absence" of a contiguous patch of similar pixels (a super-pixel), while the classifier may represent the image as a tensor with three color channels per pixel. We denote $x \in \mathbb{R}^d$ be the original representation of an instance being explained, and we use $x' \in \{0,1\}^{d'}$ to denote a binary vector for its interpretable representation.

***First Term**: the measure of the unfaithfulness of g in approximating f in the locality defined by Pi. This is termed as **locality-aware loss** in the original paper*

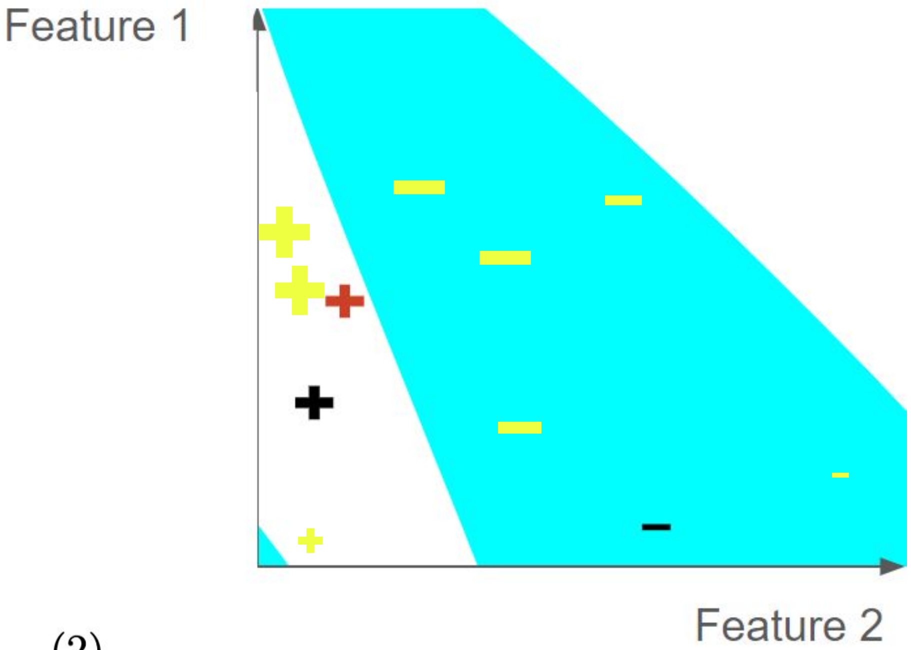
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \underline{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$

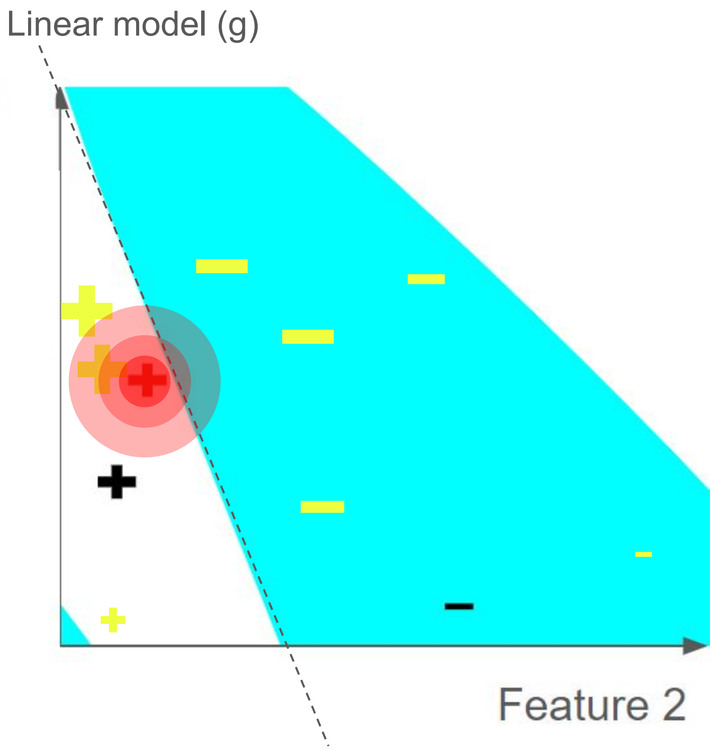
Weighted on the distance of z to x

Label complex model

Prediction simple model



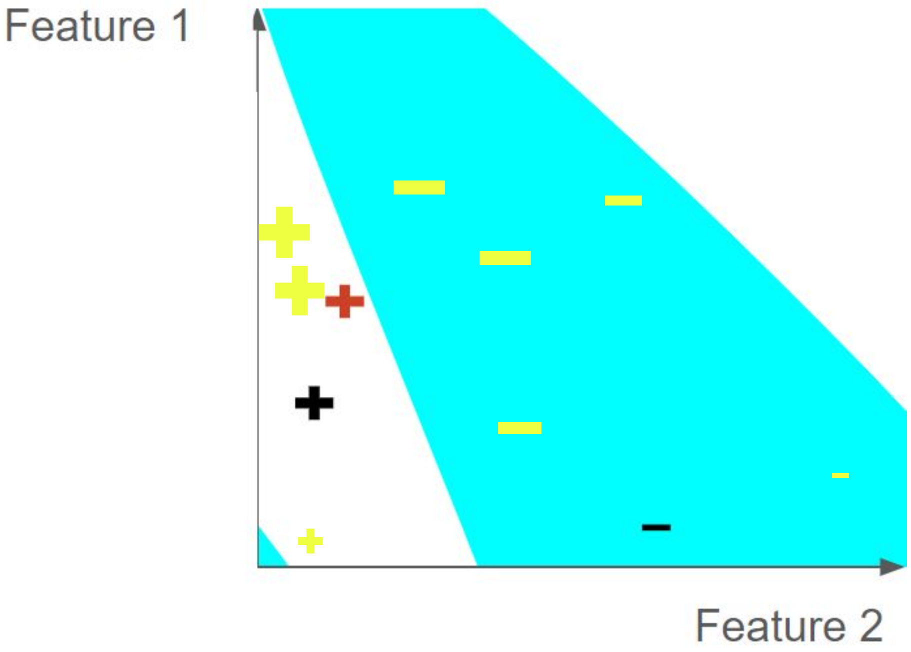
$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$



3.4 Sparse Linear Explanations

For the rest of this paper, we let G be the class of linear models, such that $g(z') = w_g \cdot z'$. We use the locally weighted square loss as $\mathcal{L}$, as defined in Eq. ($\boxed{2}$), where we let $\pi_x(z) = exp(-D(x, z)^2/\sigma^2)$ be an exponential kernel defined on some

Last term: *a measure of model complexity of explanation g. For example, if your explanation model is a decision tree it can be the depth of the tree or in the case of linear explanation models it can be the number of non zero weights*



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \underline{\Omega(g)}$$

For text classification, we ensure that the explanation is **interpretable** by letting the *interpretable representation* be a bag of words, and by setting a limit K on the number of words, i.e. $\Omega(g) = \infty \mathbb{1}[\|w_g\|_0 > K]$. Potentially, K can be