Robust Machine Learning

Sherry Ruan

October 19, 2015

In reality, training data distribution and test data distribution could be quite different from each other. This complicates the construction of machine learning models because we want to train classifiers robust against such changes. Different approaches have been proposed to tackle the problem. We summarize and compare three methods here:

Robust optimization In this work, the authors proposed to build classifiers by a robust minimax approach. Their classifier is robust against the test time deletion of input data and optimal in the worst case deletion situation (i.e., minimize the worst case hinge loss:

$$\min_{w\in\mathbb{R}^d}\sum_{i=1}^{n}\max_{\left\{\alpha_i\in\{0,1\}^d:\sum_{j=1}^{n}\alpha_{ij}=n-K\right\}}\lfloor 1-y_i w\cdot x_i\circ\alpha_i\rfloor_+$$

where x_i are original examples and α_i is a vector which has K zero elements corresponding to the deletion of K features in x_i. The authors showed that the minimization of the hinge loss could be transformed to a quadratic programming problem and hence effectively solvable.

Covariate shift The main assumption in this approach is that the conditional probability distribution of the output is fixed(same?) in the training and test set, expressed mathematically as $P_{tr}(d_y|x) = P_{te}(d_y|x)$. Under such an assumption, we could solve the problem (estimate the expected value of the test output) by an estimation from the training set. The core estimation is the Radon-Nikodym derivative $\beta(x) = \frac{dP_{te}(x)}{dP_{tr}}$. The key assumption here is that $\beta(x)$ is well-defined, i.e., there does not exist any measurable set A such that $P_{te}(A) > 0$ and $P_{tr}(A) = 0$. The authors proposed a kernel

mean matching (KMM) algorithm to transform the estimation problem to a quadratic programming problem.

Well specification Recall $P(x, y) = P(y|x)P(x)$, so provided that $P(y|x)$ is given, we can train a classifier which is robust against other input distribution $P(x)$. In case that $P(y|x)$ is not defined over the entire input space, we would have preconditions. Note that in both cases we should verify that $P(x)$ is not degenerating in its domain.

Comparison These three methods all tried to address the problem when the true data distribution is hard to obtain. They tackle the problem by making different assumptions. *Robust optimization* assumes a worst case data perturbation/deletion scenario. *Covariate shift* relies on the assumption that the train data distribution is intrinsically related to the test data distribution, and *well specification* depends on the specification of $y|x$ and the existence of $p(x)$. Both *robust optimization* and *covariate shift* solved the problem effectively by transforming the minimization/estimation problem to a quadratic problem.

Types of scenarios There are a variety of real world scenarios which call for a robust model against the loss of the data. For example, in unreliable communication channels, we may have stochastic data impairments during the transmission. Another example is that during the compression of images to JPEG forms, we may lose a certain portion of information of images. We may also encounter constant changes of data (in a neighborhood) such as the recognition of moving objects in robotics. Also consider an online recommendation system where recommendations are largely dependent on the user's previous choices, which in turn depend on the recommendations. It may be uncommon to have a real adversary which always corrupts the robustness of our models as much as possible, but think of a rival which may deliberately insert malicious data because of some economic incentives or some virus which constantly generate fake information.

Thinking: Differences between the distributional view and the max view When we consider the perturbation of the data, we can either consider features are changed based on some probability distribution or we may assume that features are altered/dropped in a worst case. In general,

the later case is more pessimistic but less common in reality. We can regard the max view as a special discrete probability distribution where all density concentrates at one point.

Problem Formulation We set up the problem as follows: Given n examples (x_i, y_i) where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, we want to build a linear model which is robust against the perturbation of the data. More specifically, we want to find $w \in \mathbb{R}^d$ which minimizes the hinge loss function (and the squared loss function) over the following distributions:

1 Worst case in a closed ball with radius r

1.1 Squared loss

Now we consider the scenario of squared loss functions. We still assume that for each data point x_i, the perturbation $\tilde{x_i}$ is generated from a closed ball with radius r centered at x_i. We consider the worst case squared loss function here.

$$w^* = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_{\{\tilde{x_i} : \|\tilde{x_i} - x_i\| \leq r\}} (y_i - w \cdot \tilde{x_i})^2 \right\}$$

1.1.1 L1-norm

We consider the L1 - norm.

$$w^* = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_{\{\tilde{x_i} : \|\tilde{x_i} - x_i\|_1 \leq r\}} (y_i - w \cdot \tilde{x_i})^2 \right\}$$

Let $\alpha_i = x_i - \tilde{x_i}$ and we consider the inner max function first:

$$\max_{\|\alpha_i\|_1 \leq r} [(y_i - w \cdot x_i) + (w \cdot \alpha_i)]^2$$

Since $[(y_i - w \cdot x_i) + (w \cdot \alpha_i)]^2$ is a quadratic function, we want to maximize $|w \cdot \alpha_i|$ as the max will be achieved at either of the endpoints. Without loss of generality, suppose $|w_k| = \max_{j=1}^{d} |w_j| = \|w\|_\infty$

$$|w \cdot \alpha_i| = \left| \sum_{j=1}^{d} w_j \alpha_{ij} \right| \leq \sum_{j=1}^{d} |w_j| \, |\alpha_{ij}| \leq |w_k| \sum_{j=1}^{d} |\alpha_{ij}| \leq \|w\|_\infty \, r$$

The equality holds when $\alpha_i = re_k$. Therefore, the original problem becomes:

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(|y_i - w\cdot x_i| + \|w\|_\infty r)^2 \right\}$$

1.1.2 L2-norm

We consider the L2 - norm.

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max_{\{\tilde{x_i}:\|\tilde{x_i}-x_i\|_2\le r\}} (y_i - w\cdot\tilde{x_i})^2 \right\}$$

Let $\alpha_i = x_i - \tilde{x_i}$ and we consider the inner max function first:

$$\max_{\|\alpha_i\|_2\le r} [(y_i - w\cdot x_i) + (w\cdot\alpha_i)]^2$$

Note that by Cauchy-Schwartz inequality, $|w\cdot\alpha_i| \le \|w\|_2 \|\alpha_i\|_2 \le r\|w\|_2$. Because $[(y_i - w\cdot x_i) + (w\cdot\alpha_i)]^2$ is quadratic, the max is achieved at $|w\cdot\alpha_i| = r\|w\|_2$.

Let $\alpha_i = \frac{\pm r}{\|w\|_2} w$, we can get $w\cdot\alpha_i = \pm r\|w\|_2$ and the max is $(y_i - w\cdot x_i)^2 + r^2\|w\|_2^2 + 2r|y_i - w\cdot x_i|\,\|w\|_2$. Hence, the original minimization problem becomes:

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \left((y_i - w\cdot x_i)^2 + r^2\|w\|^2 + 2r|y_i - w\cdot x_i|\,\|w\| \right) \right\}$$

$$= \min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} (|y_i - w\cdot x_i| + r\|w\|)^2 \right\}$$

This can be viewed as adding some penalty on the weights to the original objective function. But note that it depends on the difference between predicted labels and given labels that how much penalty we want to add.

Suppose we add an addition constraint that $\|w\| = 1$, then the objective function above can be simplified to

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2} + C\sum_{i=1}^{n} (y_i - w\cdot x_i)^2 + r^2 + 2r|y_i - w\cdot x_i| \right\}$$

$$= \min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} + C \sum_{i=1}^{n} (|y_i - w \cdot x_i| + r)^2 \right\}$$

Which can be regarded as adding some noise r to the squared loss function.

1.1.3 Infinity norm

We consider the L∞ - norm.

$$w^* = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_{\{\tilde{x_i} : \|\tilde{x_i} - x_i\|_\infty \leq r\}} (y_i - w \cdot \tilde{x_i})^2 \right\}$$

Let $\alpha_i = x_i - \tilde{x_i}$ and we consider the inner max function first:

$$\max_{\|\alpha_i\|_\infty \leq r} [(y_i - w \cdot x_i) + (w \cdot \alpha_i)]^2$$

Since $[(y_i - w \cdot x_i) + (w \cdot \alpha_i)]^2$ is a quadratic function, we want to maximize $|w \cdot \alpha_i|$ as the max will be achieved at either of the endpoints. Without loss of generality, suppose $|\alpha_{ik}| = \max_{j=1}^{d} |\alpha_{ij}| = \|\alpha_i\|_\infty$

$$|w \cdot \alpha_i| = \left| \sum_{j=1}^{d} w_j \alpha_{ij} \right| \leq \sum_{j=1}^{d} |w_j| \, |\alpha_{ij}| \leq |\alpha_{ik}| \sum_{j=1}^{d} |w_j| = \|\alpha_i\|_\infty \|w\|_1 \leq r \|w\|_1$$

The equality holds when $\alpha_i = r \sum_{j=1}^{d} e_j$. Therefore, the original problem becomes:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (|y_i - w \cdot x_i| + \|w\|_1 r)^2 \right\}$$

1.2 Hinge loss function

We assume that for each data point x_i, the perturbation $\tilde{x_i}$ is generated from a closed ball with radius r centered at x_i. We consider the hinge loss function in the worst case scenario here.

$$w^* = \arg\min_{w \in \mathbb{R}^d} \{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max_{\{\tilde{x_i} : \|\tilde{x_i} - x_i\| \leq r\}} (1 - y_i w \cdot \tilde{x_i})_+ \}$$

Note that we take account of a regularization term to penalize the weight vector norm (C is a constant chosen)

First, let $\alpha_i = x_i - \tilde{x_i}$, then the constraint $\|\tilde{x_i} - x_i\| \leq r$ becomes $\|\alpha_i\| \leq r$. We can re-formulate the problem as follows:

$$\max_{\alpha_i \in \mathbb{R}^d} \ (1 - y_i w \cdot (x_i - \alpha_i))_+ \qquad s.t. \|\alpha_i\| \leq r \qquad \text{for } i = 1, \cdots, n$$

In order to maximize this function, we want $y_i w \cdot \alpha_i$ to be as large as possible, so the problem can be further simplified to:

$$\max_{\alpha_i \in \mathbb{R}^d} \ y_i w \cdot \alpha_i \qquad s.t. \|\alpha_i\| \leq r \qquad \text{for } i = 1, \cdots, n$$

1.2.1 L1-norm

We consider the 1-norm here. By writing out the elements of α_i explicitly, we get:

$$\max \ y_i(\sum_{j=1}^{d} w_j \alpha_{ij}) \qquad s.t. \sum_{j=1}^{d} |\alpha_{ij}| \leq r$$

We introduce d variables b_{ij}, then the problem is equivalent to:

$$\max \ y_i(\sum_{j=1}^{d} w_j \alpha_{ij}) \qquad s.t. \sum_{j=1}^{d} b_{ij} \leq r \qquad |\alpha_{ij}| \leq b_{ij} \quad \forall j \in \{1, \cdots, d\}$$

We rewrite it in a neat way:

$$\max \ y_i(\sum_{j=1}^{d} w_j \alpha_{ij}) \qquad s.t. \sum_{j=1}^{d} b_{ij} \leq r$$

$$\alpha_{ij} + b_{ij} \geq 0 \quad \forall j \in \{1, \cdots, d\}$$

$$\alpha_{ij} - b_{ij} \leq 0 \quad \forall j \in \{1, \cdots, d\}$$

$$b_{ij} \geq 0, \quad \alpha_{ij} \text{ free} \quad \forall j \in \{1, \cdots, d\}$$

We introduce a dual variable c_i to represent the first inequality and $z_{i1} \cdots z_{id}$ to represent $\alpha_{ij} + b_{ij} \geq 0$ and $s_{i1} \cdots s_{id}$ to represent $\alpha_{ij} - b_{ij} \leq 0$. Then the dual problem is:

$$\min \ r c_i$$

$$s.t.\ \ c_i \geq 0$$

$$z_{ij} + s_{ij} = y_i w_j \qquad \text{for } j = 1, \cdots, d$$

$$z_{ij} - s_{ij} + c_i \geq 0 \qquad \text{for } j = 1, \cdots, d$$

$$z_{ij} \leq 0, \quad s_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d$$

We then plug this dual problem into the original minimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (1 + t_i - y_i w \cdot x_i)_+$$

$$s.t.\ \ t_i \geq\ rc_i$$

$$c_i \geq 0$$

$$z_{ij} + s_{ij} = y_i w_j \qquad \text{for } j = 1, \cdots, d$$

$$z_{ij} - s_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d$$

$$z_{ij} \leq 0, \quad s_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d$$

This is a quadratic problem with linear constraints, hence we can solve it efficiently.

1.2.2 L2-norm

Suppose we only have one dimension, then for each i, the maximum is achieved when $\tilde{x}_i$ is $x_i \pm r$ where the sign is chosen to make $y_i w r$ positive, so the problem is also convex hence efficiently solvable:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (1 - y_i w x_i + |y_i w r|)_+$$

The same idea applies to d-dimension. Here we consider the 2-norm function instead of the 1-norm in the constraints. Then we are maximizing a linear function over a convex set. The maximum should be achieved at the boundary.

$$\max_{\alpha_i \in \mathbb{R}^d} y_i w \cdot \alpha_i \qquad s.t. \|\alpha_i\|_2 = r \qquad \text{for } i = 1, \cdots, n$$

Note that by Cauchy-Schwartz inequality, $|w \cdot \alpha_i| \leq \|w\|_2 \|\alpha_i\|_2 = r \|w\|_2$

One feasible assignment is $\alpha_i = \frac{\pm r}{\|w\|_2} w$ (because $\|\alpha_i\|_2 = r$), then we can get $y_i w \cdot \alpha_i = y_i w \cdot \frac{\pm r}{\|w\|_2} w = \pm y_i r \|w\|_2$. The sign is chosen to make $y_i r$ positive. This is exactly when the max of the function is achieved. Hence, the original minimization problem is transformed to an easy quadratic problem:

$$\min_{w \in \mathbb{R}^d} \{\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (1 - y_i w \cdot x_i + |y_i r| \|w\|_2)_+\}$$

where the worst case $\tilde{x_i}$ is chosen as (sign depending on y_i):

$$\tilde{x_i} = x_i \pm \frac{r}{\|w\|_2} w$$

Note that $y_i \in \{\pm 1\}$. Suppose $r = 1$, the problem can be further simplified to

$$\min_{w \in \mathbb{R}^d} \{\frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} ((1 + \|w\|_2) - y_i w \cdot x_i)_+\}$$

Note by Laplacian transformation, there exists r' such that the formula above is equivalent to

$$\min_{w \in \mathbb{R}^d} \{\sum_{i=1}^{n} ((1 + \|w\|_2) - y_i w \cdot x_i)_+\} \qquad \text{where } \|w\| \leq r'$$

If we assume $\|w\| = r'$, then we just get

$$\min_{w \in \mathbb{R}^d} \{\sum_{i=1}^{n} ((1 + r') - y_i w \cdot x_i)_+\}$$

This can actually be viewed as a hinge loss problem with a different boundary.

1.2.3 Infinity norm

We also consider the infinity norm here. By replacing each absolute value with two inequalities, we get:

$$\max \ y_i \left(\sum_{j=1}^{d} w_j \alpha_{ij} \right) \qquad s.t. \alpha_{ij} \geq -r, \alpha_{ij} \leq r \quad \forall j \in \{1, \cdots, d\}$$

We introduce dual variables $z_{i1} \cdots z_{id}$ to represent d inequalities $\alpha_{ij} \geq -r$, and dual variables $s_{i1} \cdots s_{id}$ to represent d inequalities $\alpha_{ij} \leq r$ so the dual problem is:

$$\begin{aligned} \min\ & r\left(\sum_{j=1}^{d} s_{ij}\right) - r\left(\sum_{j=1}^{d} z_{ij}\right) \\ s.t.\ \ & z_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d \\ & s_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d \\ & z_{ij} + s_{ij} = y_i w_j \qquad \text{for } j = 1, \cdots, d \end{aligned}$$

We then plug this dual problem into the original minimization problem:

$$\begin{aligned} \min_{w \in \mathbb{R}^d}\ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(1 + t_i - y_i w \cdot x_i)_+ \\ s.t.\ \ & t_i \geq r\left(\sum_{j=1}^{d} s_{ij}\right) - r\left(\sum_{j=1}^{d} z_{ij}\right) \\ & z_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d \\ & s_{ij} \geq 0 \qquad \text{for } j = 1, \cdots, d \\ & z_{ij} + s_{ij} = y_i w_j \qquad \text{for } j = 1, \cdots, d \end{aligned}$$

Hence we transformed the original problem to a quadratic problem. By observing the structure of the transformed objective function, we notice that it is also a hinge loss function with a different boundary.

2 Noise generated from a Gaussian distribution

2.1 Squared loss function

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left\{\sum_{i=1}^{n}(y_i - w \cdot \tilde{x_i})^2\right\} \quad \text{where } \tilde{x_i} \sim \mathcal{N}_d(x_i, \Sigma)$$

Note that for $X \sim \mathcal{N}(\mu, \Sigma)$, $Y = C + BX$, we have $Y \sim \mathcal{N}(C + B\mu, B\Sigma B^T)$, so we have:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^{n}(y_i - w \cdot \tilde{x_i})^2\right\} &= \sum_{i=1}^{n}\mathbb{E}\left\{(y_i - w \cdot \tilde{x_i})^2\right\} \\
&= \sum_{i=1}^{n}\mathbb{E}\left\{y_i^2 + (w \cdot \tilde{x_i})^2 - 2y_i(w \cdot \tilde{x_i})\right\} \\
&= \sum_{i=1}^{n}\left\{y_i^2 + \mathbb{E}((w \cdot \tilde{x_i})^2) - 2y_i\mathbb{E}(w \cdot \tilde{x_i})\right\} \\
&= \sum_{i=1}^{n}\left\{y_i^2 + (w \cdot x_i)^2 + (w\Sigma w^T) - 2y_i w \cdot x_i\right\}
\end{aligned}$$

Without loss of generality, we assume Σ is the identity matrix then the equation above can be simplified to:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^{n}(y_i - w \cdot \tilde{x_i})^2\right\} &= \sum_{i=1}^{n}\left\{y_i^2 + (w \cdot x_i)^2 + \|w\|^2 - 2y_i w \cdot x_i\right\} \\
&= \sum_{i=1}^{n}\left\{y_i^2 + (w \cdot x_i)^2 - 2y_i w \cdot x_i\right\} + n\|w\|^2 \\
&= \left(\sum_{i=1}^{n}\left(y_i - w \cdot x_i\right)^2\right) + n\|w\|^2
\end{aligned}$$

From the observation of the equation above, we realize that minimizing the squared loss function with Gaussian noise is equivalent to minimizing the original squared loss function plus a penalty on the weight which is linear in n.

2.2 Hinge loss function

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left\{\sum_{i=1}^{n}(1 - y_i w \cdot \tilde{x_i})_+\right\} \quad \text{where } \tilde{x_i} \sim \mathcal{N}(x_i, \Sigma)$$

We consider the expectation part first:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^{n}(1-y_i w\cdot\tilde{x_i})_+\right\} &= \sum_{i=1}^{n}\mathbb{E}\left\{\max(1-y_i w\cdot\tilde{x_i},0)\right\} \\
&= \sum_{i=1}^{n}\{\mathbb{E}(1-y_i w\cdot\tilde{x_i}|1-y_i w\cdot\tilde{x_i}>0)\mathbb{P}(1-y_i w\cdot\tilde{x_i}>0) \\
&\quad +\mathbb{E}(0|1-y_i w\cdot\tilde{x_i}\leq 0)\mathbb{P}(1-y_i w\cdot\tilde{x_i}\leq 0)\} \\
&= \sum_{i=1}^{n}\mathbb{E}(1-y_i w\cdot\tilde{x_i}|1-y_i w\cdot\tilde{x_i}>0)\mathbb{P}(1-y_i w\cdot\tilde{x_i}>0)
\end{aligned}$$

Note that since $\tilde{x_i}\sim\mathcal{N}(x_i,\Sigma)$, we have $(1-y_i w\cdot\tilde{x_i})\sim\mathcal{N}(1-y_i w x_i,(y_i w)\Sigma(y_i w)^T)$. Hence,

$$\mathbb{E}(1-y_i w\cdot\tilde{x_i}|1-y_i w\cdot\tilde{x_i}>0)=\mathbb{E}(|1-y_i w\cdot\tilde{x_i}|)$$

is actually a **folded normal distribution**. The mean is just

$$\sigma\sqrt{\frac{2}{\pi}}e^{\frac{-\mu^2}{2\sigma^2}}+\mu\left(1-2\Phi(\frac{-\mu}{\sigma})\right)$$

Where Φ is the CDF of the standard normal distribution, $\sigma=\sqrt{(y_i w)\Sigma(y_i w)^T}$, and $\mu=1-y_i w x_i$.

The other part we need to solve is $\mathbb{P}(1-y_i w\cdot\tilde{x_i}>0)$. This is easy because

$$\begin{aligned}
\mathbb{P}(1-y_i w\cdot\tilde{x_i}>0) &= \mathbb{P}\left(1-y_i w\cdot\tilde{x_i}-(1-y_i w\cdot x_i)>-(1-y_i w\cdot x_i)\right) \\
&= \mathbb{P}(-y_i w\cdot\tilde{x_i}+y_i w\cdot x_i)>-1+y_i w\cdot x_i) \\
&= \mathbb{P}\left(\frac{-y_i w\cdot\tilde{x_i}+y_i w\cdot x_i}{|(y_i w)\Sigma(y_i w)^T|}>\frac{-1+y_i w\cdot x_i}{|(y_i w)\Sigma(y_i w)^T|}\right) \\
&= \Phi\left(-\frac{-1+y_i w\cdot x_i}{|(y_i w)\Sigma(y_i w)^T|}\right) \\
&= \Phi\left(\frac{\mu}{\sigma}\right)
\end{aligned}$$

because the left hand side has standard normal distribution.

Hence, combining these two parts, we get

$$\mathbb{E}\left\{\sum_{i=1}^{n}(1-y_i w\cdot\tilde{x_i})_+\right\}=\sum_{i=1}^{n}\left\{\sigma\sqrt{\frac{2}{\pi}}e^{\frac{-\mu^2}{2\sigma^2}}+\mu\left(1-2\Phi\left(\frac{-\mu}{\sigma}\right)\right)\Phi\left(\frac{\mu}{\sigma}\right)\right\}$$

3 Worst case dropout

Assuming K features get deleted

$$\min_{w\in\mathbb{R}^d}\sum_{i=1}^{n}\max_{\left\{\alpha_i\in\{0,1\}^d:\sum_{j=1}^{n}\alpha_{ij}=n-K\right\}}(y_i-w\cdot x_i\circ\alpha_i)^2$$

$$\min_{w\in\mathbb{R}^d}\sum_{i=1}^{n}\max_{\left\{\alpha_i\in\{0,1\}^d:\sum_{j=1}^{n}\alpha_{ij}=n-K\right\}}(1-y_iw\cdot x_i\circ\alpha_i)_+$$

The solution to the problem is described in the nightmare paper.

4 Dropout noise distribution

$$\min_{w\in\mathbb{R}^d}\mathbb{E}\left\{\sum_{i=1}^{n}(y_i-w\cdot(x_i\circ\delta_i))^2\right\}\quad\text{where }\delta_i\in\left\{0,\frac{1}{1-\delta}\right\}^d\text{ are independent Bernoulli r.v.s}$$

$$\min_{w\in\mathbb{R}^d}\mathbb{E}\left\{\sum_{i=1}^{n}(1-y_iw\cdot(x_i\circ\delta_i))_+\right\}\quad\text{where }\delta_i\in\left\{0,\frac{1}{1-\delta}\right\}^d\text{ are independent Bernoulli r.v.s}$$

The solution to the problem is described in the dropout paper.

5 Summary

5.1 Squared Loss

Squared loss + L1 noise: Penalty on $\|w\|_\infty$

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(|y_i - w\cdot x_i| + r\|w\|_\infty)^2 \right\}$$

Squared loss + L2 noise: Add r to the squared loss

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(|y_i - w\cdot x_i| + r\|w\|_2)^2 \right\}$$

Squared loss + L-∞ noise: Penalty on $\|w\|_1$

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(|y_i - w\cdot x_i| + r\|w\|_1)^2 \right\}$$

In general, if we have L-p noise, then the corresponding objective function will be:

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(|y_i - w\cdot x_i| + r\|w\|_q)^2 \right\}$$

where q is the conjugate index of p.

Squared loss + Gaussian noise: Equivalent to L2 regularization with a penalty on weights linear in n

$$n\|w\|^2 + \sum_{i=1}^{n}(y_i - w\cdot x_i)^2$$

5.2 Hinge Loss

Hinge loss + L1 noise: Increase the margin threshold by t_i

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}((1+t_i) - y_i w\cdot x_i)_+$$

Hinge loss + L-2 noise: : Increase the margin threshold by $\|w\|$

$$\min_{w\in\mathbb{R}^d} \{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}((1+|r|\,\|w\|) - y_i w\cdot x_i)_+\}$$

Suppose $r=1$ and $\|w\|=1$, then we get

$$\min_{w\in\mathbb{R}^d} \{\frac{1}{2} + C\sum_{i=1}^{n}(2 - y_i w\cdot x_i)_+\}$$

Hinge loss + L-∞ noise: : Increase the margin threshold by t_i

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}((1+t_i) - y_i w\cdot x_i)_+$$

Hinge loss + Gaussian noise: Too complex!

$$\mathbb{E}\left\{\sum_{i=1}^{n}(1-y_i w\cdot \tilde{x_i})_+\right\} = \sum_{i=1}^{n}\left\{\sigma\sqrt{\frac{2}{\pi}}e^{\frac{-\mu^2}{2\sigma^2}} + \mu\left(1-2\Phi\left(\frac{-\mu}{\sigma}\right)\right)\Phi\left(\frac{\mu}{\sigma}\right)\right\}$$