The Australian National University Semester 1, 2022
School of Computing Assignment 2
Group member name(s): Campbell Rankine
Group member UID(s): u6997593

COMP4670: Statistical Machine Learning

1 Section 1: Generalization and Data Noise

Answer to Question 1.1: Noise and True Risk

By definition of True and Empirical risk:

$$|R_D[f] - R_S[f]| = \int\int |\mathbb{E}_{(x,y)\sim D}[\mathcal{L}(f(x), y)] - \mathbb{E}_{(x,y)\sim S}[\mathcal{L}(f(x), y)]|dxdy$$

$$|R_D[f] - R_S[f]| = \int\int |p(x,y)\mathcal{L}(f(x), y) - \tilde{p}(x,y)\mathcal{L}(f(x), y)|dxdy$$

$$|R_D[f] - R_S[f]| = \int\int |(p(x,y) - \tilde{p}(x,y))(\mathcal{L}(f(x), y)|dxdy$$

Since we know the loss function is bounded by [0,1], we can relax the bounds and set the value of the loss to its max value:

$$|R_D[f] - R_S[f]| \leq \int\int |p(x,y) - \tilde{p}(x,y)|dxdy$$

$$|R_D[f] - R_S[f]| \leq D_{tv}(p, \tilde{p})$$

Answer to Question 1.2: Uniform bound for the noisy gap

By introducing the triangle inequality to expand the term we are seaking to upper bound, we get:

$$|R_D[f] - R_{\tilde{D}}[f]| + |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \geq |R_D[f] - R_{\tilde{S}}[f]|$$

From question 1 we can upper bound the left term in the LHS with:

$$D_{tv}(p, \tilde{p}) + |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \geq |R_D[f] - R_{\tilde{S}}[f]|$$

Now since we want to find an upper bound of the RHS with probability $1 - \delta$, one method we can use to do this is find the upper bound of $|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|$ with probability $1 - \delta$. Doing so will yield us the result we are looking for. If we can successfully do this, we can also prove that the standard generalization gap is indiscriminant to the noise.

Start by applying Hoeffding's:

$$P(|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \epsilon) \geq 1 - 2Exp(-2N\epsilon^2), \forall f \epsilon H$$

Find Delta:

$$\delta = 2Exp(-2N\epsilon^2)$$
$$\delta/2 = Exp(-2N\epsilon^2)$$

$$ln(\delta/2) = -2N\epsilon^2$$

$$\frac{-ln(\delta/2)}{2N} = \epsilon^2$$

$$\frac{ln(2/\delta)}{2N} = \epsilon^2$$

$$\sqrt{\frac{ln(2/\delta)}{2N}} = \epsilon$$

Introduce hoefdings: '

$$|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \sqrt{\frac{ln(2/\delta)}{2N}}$$

Relax bounds:

$$|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \sqrt{\frac{ln(1/\delta)}{N}}$$

We are looking for a bound with complexity constraint using $\forall f \epsilon H$ and we know:

$$P(|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \epsilon) = P(\exists f \epsilon H : |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|)$$

$$P(\exists f \epsilon H : |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|) \leq \sum_{f \epsilon H} P(|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \epsilon)$$

$$P(\exists f \epsilon H : |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|) \leq \sum_{f \epsilon H} \delta$$

$$P(\exists f \epsilon H : |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|) \leq |H|\delta$$

Therefore:

$$|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \sqrt{\frac{ln((1/\delta)|H|)}{N}}$$

$$|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}$$

With probability $1 - \delta$, $\forall f \epsilon H$

Since we have now bounded $|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]|$ with probability $1 - \delta$:

$$D_{tv}(p, \tilde{p}) + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}} \geq |R_D[f] - R_{\tilde{S}}[f]|$$

With probability $1 - \delta$, $\forall f \epsilon H$

Therefore:

$$P(|R_D[f] - R_{\tilde{S}}[f]| \leq D_{tv}(p, \tilde{p}) + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}) \geq 1 - \delta$$

Answer to Question 1.3: Interpreting the Uniform Bound for Noisy Sample Gap

Given our two equations:

$$P(|R_D[f] - R_S[f]| \leq \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}) \geq 1 - \delta$$

$$P(|R_D[f] - R_{\tilde{S}}[f]| \leq D_{tv}(p, \tilde{p}) + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}) \geq 1 - \delta$$

We can see that our upper bounds with probability $1 - \delta$ are given by:

$$|R_D[f] - R_S[f]| \leq \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}$$

and

$$|R_D[f] - R_{\tilde{S}}[f]| \leq D_{tv}(p, \tilde{p}) + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}$$

Since we consider maximum possible error for our bound for comparison between the two, we can definitively say that:

$$|R_D[f] - R_S[f]| \leq |R_D[f] - R_{\tilde{S}}[f]|$$

Where the two are only equal when the noisy sample = true sample. Meaning that sampling noisy data increases the generalization gap, which would normally be considered a bad thing.

For space complexity we know that the space complexity of $|R_D[f] - R_S[f]|$ is: (From lectures given to us)

$$O(\sqrt{\frac{complexity}{N}})$$

For our noisy sample gap:

$$O(D_{tv}(p, \tilde{p}) + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}})$$

$$O(C + \sqrt{\frac{ln(|H|)}{N}})$$

since D_{tv} and $ln(1/\delta)$ do not rely on N. When considering a C in big O complexity we generally drop this value so:

$$O(\sqrt{\frac{ln(|H|)}{N}})$$

$$O(\sqrt{\frac{complexity}{N}})$$

So while a noisy sample gap increases the overall generalization gap, it does not increase the space complexity of our noisy sample gap.

Answer to Question 1.4: Noisy Posterior for Label Noise

Using composition of conditional probabilities we can arrive at the formula:

$$P(\tilde{Y} = y \mid X) = P(\tilde{Y} = y \mid Y = y)P(Y = -y \mid X) + P(\tilde{Y} = -y \mid Y = y)P(Y = y \mid X)$$

The opening probabilites on the RHS are seen in the definition of Asymmetric label noise LN1:

$$P(\tilde{Y} = y \mid X) = \sigma_{-y}P(Y = -y \mid X) + (1 - \sigma_y)P(Y = y \mid X)$$

Put everything in terms of $Y = y$

$$P(\tilde{Y} = y \mid X) = \sigma_{-y}(1 - P(Y = y \mid X)) + (1 - \sigma_y)P(Y = y \mid X)$$

Answer to Question 1.5: Uniform Bound for Symmetric Label Noise

Start by bounding $|R_D[f] - R_{\tilde{D}}[f]|$:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq D_{tv}(p, \tilde{p})$$
$$D \sim p(x, y)$$
$$\tilde{D} \sim p(\tilde{x}, \tilde{y})$$

Using definition of D_{tv}:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(x, y) - p(\tilde{x}, \tilde{y})|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x)p(x) - p(\tilde{y} \mid \tilde{x})p(\tilde{x})|dxdy$$

By LN2 (Identical input margins):

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x)p(x) - p(\tilde{y} \mid \tilde{x})p(x)|dxdy$$
$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(x)(p(y \mid x) - p(\tilde{y} \mid \tilde{x}))|dxdy$$

since $p(x)$ has a max value of 1, relax bounds by setting it to 1:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x) - p(\tilde{y} \mid \tilde{x})|dxdy$$

Examine $p(\tilde{y} \mid \tilde{x})$:

$$p(\tilde{y} \mid \tilde{x}) = \frac{p(\tilde{x} \mid \tilde{y})p(\tilde{y})}{p(\tilde{x})}$$

since LN2 says that $p(x) = p(\tilde{x})$ and also implies that $p(\tilde{x} \mid \tilde{y}) = p(x \mid \tilde{y})$ in our original equation we can manipulate this to:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x) - p(\tilde{y} \mid x)|dxdy$$

Using equation 1.8 and the definition of Symmetric label noise we obtain:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x) - (1-2\sigma)p(y \mid x) + \sigma|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x)(1-(1-2\sigma)) + \sigma|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |p(y \mid x)(-2\sigma) + \sigma|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |(-2\sigma)(p(y \mid x) + -0.5)|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq \int\int |(-2\sigma)||(p(y \mid x) + -0.5)|dxdy$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq |(-2\sigma)|\int\int |(p(y \mid x) + -0.5)|dxdy$$

We know D_{tv} can be converted to a summation formula so:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq |(-2\sigma)|\sum_X \sum_Y |(p(y \mid x) + -0.5)|$$

Set $p(y \mid x)$ to its max value 1 for all values:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq |(-2\sigma)|2(|(1 + -0.5)|)$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq |(-2\sigma)|2(0.5)$$

Arriving us at our final result:

$$|R_D[f] - R_{\tilde{D}}[f]| \leq |(-2\sigma)|$$

$$|R_D[f] - R_{\tilde{D}}[f]| \leq 2\sigma$$

As before introduce triangle inequality:

$$|R_D[f] - R_{\tilde{D}}[f]| + |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \geq |R_D[f] - R_{\tilde{S}}[f]|$$

$$2\sigma + |R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \geq |R_D[f] - R_{\tilde{S}}[f]|$$

In question 1.2 we prove that the bound for our uniform gap is indiscriminant of noise, therefore we can use the same result that:

$$|R_{\tilde{D}}[f] - R_{\tilde{S}}[f]| \leq \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}$$

With probability $1-\delta$, $\forall f \epsilon H$

And once again with similar logic to question 1.2 we can conclude that:

$$P(|R_D[f] - R_{\tilde{S}}[f]| \leq 2\sigma + \sqrt{\frac{ln(|H|) + ln(1/\delta)}{N}}) \geq 1-\delta$$

2 Section 2: Gaussian Process and Bayesian Optimization

Answer to Question 2.3: Derivation of Expected Improvement

From Equation 2.8:

$$I(x) = max(0, f(x) - f^{+}(x) - \epsilon)$$

$$E[I(x)] = \int f(x) - -f^{+}(x) - \epsilon df$$

Since $f(x)$ can be expressed by: (We know its a gaussian)

$$f(x) \sim \mu(x) + \sigma(x)\mathcal{N}(0,1)$$

$$E[I(x)] = \int (\mu(x) + \sigma(x) - f^{+}(x) - \epsilon)(\phi(x'')dx''$$

$$E[I(x)] = \int_{-inf}^{Z} (\mu(x) + \sigma(x)x'' - f^{+}(x) - \epsilon)(\phi(x'')dx''$$

$$E[I(x)] = (\mu(x) - f^{+}(x) - \epsilon)\int_{-inf}^{Z} (\phi(x'')dx'' - \sigma(x)\int_{-inf}^{Z} x''(\phi(x'')dx''$$

using properties of CDF

$$E[I(x)] = (\mu(x) - f^{+}(x) - \epsilon)\Phi(Z) - \sigma(x)\int_{-inf}^{Z} x''(\phi(x'')dx''$$

$$E[I(x)] = (\mu(x) - f^{+}(x) - \epsilon)\Phi(Z) - \sigma(x)\frac{1}{(\sqrt{2\pi})} \mid Exp(\frac{-x''^{2}}{2}) \mid_{-inf}^{Z}$$

$$E[I(x)] = (\mu(x) - f^{+}(x) - \epsilon)\Phi(Z) - \sigma(x)\frac{1}{(\sqrt{2\pi})}exp(Z)$$

$$E[I(x)] = (\mu(x) - f^{+}(x) - \epsilon)\Phi(Z) - \sigma(x)\phi(Z)$$