LECTURE 15 - MAXIMUM LIKELIHOOD AND REGULARIZATION

1. Regularization

- **Overfitting/Overtraining**: We discussed overfitting and underfitting. Suppose you have data that you fit a model to, how do you know if you have overfit and/or underfit? *What is Cross Validation?*
- Previously we mentioned two common approaches to avoid overfitting:
  (1) More data: As you have more and more data, it becomes more and more difficult to "memorize" the data and its noise. Often, more data translates to the ability to use a more complex model and avoid overfitting. However, generally, you need exponentially more data with increases to model complexity. So, there is a limit to how much this helps. If you have a very complex model, you need a huge training data set.
  (2) Regularization: Regularization methods add a penalty term to the error function to discourage overfitting. These penalty terms encourage small values limiting the ability to overfit. These penalty terms are a way to trade-off between error and complexity.

$$E^*(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|_2^2 \tag{1}$$

$$= \frac{1}{2}\left(\mathbf{w}^T\mathbf{X}^T - \mathbf{t}^T\right)\left(\mathbf{w}^T\mathbf{X}^T - \mathbf{t}^T\right)^T + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{2}$$

- *What does each term mean/promote in the minimization? Why does the second term make sense for minimizing complexity?*
- *What happens to* $\mathbf{w}$ *with increasing model complexity (and no regularization)?*

$$\frac{\partial E^*(\mathbf{w})}{\partial \mathbf{w}} = 0 = \mathbf{X}^T\left(\mathbf{w}^T\mathbf{X}^T - \mathbf{t}^T\right)^T + \frac{\lambda}{2}2\mathbf{w} \tag{3}$$

$$0 = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} + \lambda\mathbf{w} \tag{4}$$

$$\mathbf{X}^T\mathbf{t} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} \tag{5}$$

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \tag{6}$$

- The l_2 norm penalty is common (one reason it is common: because it works so well mathematically with the least-squares error objective) and, so, has many names: shrinkage, ridge regression, weight decay

- *So, what happens when λ is increased? decreased? Can you think of a way to set λ?*
- We looked at the regularization term as a *penalty* term in the objective function. There is another way to interpret the regularization term as well. Specifically, there is a *Bayesian* interpretation.

$$\min E^*(\mathbf{w}) = \max -E^*(\mathbf{w}) \tag{7}$$

$$= \max \exp\left\{-E^*(\mathbf{w})\right\} \tag{8}$$

$$= \max \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w})-t_n\right)^2-\frac{\lambda}{2}\left\|\mathbf{w}\right\|_2^2\right\} \tag{9}$$

$$= \max \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w})-t_n\right)^2\right\}\exp\left\{-\frac{1}{2}\lambda\left\|\mathbf{w}\right\|_2^2\right\} \tag{10}$$

$$= \max \prod_{n=1}^{N}\exp\left\{-\frac{1}{2}\left(y(x_n,\mathbf{w})-t_n\right)^2\right\}\exp\left\{-\frac{1}{2}\lambda\left\|\mathbf{w}\right\|_2^2\right\} \tag{11}$$

- So, this is a maximization of the *data likelihood* with a *prior*: $p(\mathbf{X}|\mathbf{w})p(\mathbf{w})$

2. Method of Maximum Likelihood

- A *data likelihood* is how likely the data is given the parameter set
- So, if we want to maximize how likely the data is to have come from the model we fit, we should find the parameters that maximize the likelihood
- A common trick ot maximizing the likelihood is to maximize the log likelihood. Often makes the math much easier. *Why can we maximize the log likelihood instead of the likelihood and still get the same answer?*
- Consider: $\max \ln \exp\left\{-\frac{1}{2}\left(y(x_n,\mathbf{w})-t_n\right)^2\right\}$ We go back to our original objective.

3. Method of Maximum A Posteriori (MAP)

- Bayes Rule: $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$
- Consider: $p(\mathbf{w}|\mathscr{D}) = \frac{p(\mathscr{D}|\mathbf{w})p(\mathbf{w})}{p(\mathscr{D})}$, i.e., posterior $\propto$ likelihood $\times$ prior

4. Gaussian Distribution

- These are Gaussian distributions:

$$\mathscr{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \tag{12}$$

- σ^2 is the variance OR $\frac{1}{\sigma^2}$ is the *precision*
- So, as λ gets big, variance gets smaller/tighter. As λ gets small, variance gets larger/wider.

- What is the expected value of x?

$$E[x] = \int xp(x)dx \tag{13}$$

$$= \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx \tag{14}$$

- Change of variables: Let

$$y = \frac{x-\mu}{\sigma} \rightarrow x = \sigma y + \mu \tag{15}$$

$$dy = \frac{1}{\sigma}dx \rightarrow dx = \sigma dy \tag{16}$$

- Plugging this into the expectation:

$$E[x] = \int (\sigma y + \mu) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}y^2\right\} \sigma dy \tag{17}$$

$$= \int \frac{\sigma y}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\} dy + \int \frac{\mu}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\} dy \tag{18}$$

- The first term is an odd function: $f(-y) = -f(y)$ So, $E[x] = 0 + \mu = \mu$

5. Method of Maximum A Posteriori (MAP)

- Bayes Rule: $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$
- Lets look at this in terms of binary variables, e.g., Flipping a coin: $X = 1$ is "heads", $X = 2$ is "tails"
- Let μ be the probability of heads. If we know μ, then: $P(x = 1|\mu) = \mu$ and $P(x = 0|\mu) = 1 - \mu$

$$P(x|\mu) = \mu^x(1-\mu)^{1-x} = \begin{cases} \mu & \text{if } x = 1 \\ 1-\mu & \text{if } x = 0 \end{cases} \tag{19}$$

- This is called the *Bernoulli* distribution

$$E[x] = \mu \tag{20}$$

$$E\left[(x-\mu)^2\right] = \mu(1-\mu) \tag{21}$$

- So, suppose we conducted many Bernoulli trials (e.g., flip a coin) and we want to estimate μ
- **Method: Maximum Likelihood**

$$p(\mathscr{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) \tag{22}$$

$$= \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \tag{23}$$

– Maximize : (*What trick should we use?*)

$$\mathscr{L} = \sum_{n=1}^{N} x_n \ln \mu + (1 - x_n) \ln(1 - \mu) \tag{24}$$

$$\frac{\partial \mathscr{L}}{\partial \mu} = 0 \quad = \quad \frac{1}{\mu} \sum_{n=1}^{N} x_n - \frac{1}{1-\mu} \sum_{n=1}^{N} (1 - x_n) \tag{25}$$

$$0 \quad = \quad \frac{(1-\mu) \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} (1 - x_n)}{\mu(1-\mu)} \tag{26}$$

$$0 \quad = \quad \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} x_n - \mu \sum_{n=1}^{N} 1 + \mu \sum_{n=1}^{N} x_n \tag{27}$$

$$0 \quad = \quad \sum_{n=1}^{N} x_n - \mu N \tag{28}$$

$$\mu \quad = \quad \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N} \tag{29}$$

where m is the number of successful trials.

– So, if we flip a coin 1 time and get heads, then $\mu = 1$ and probability of getting tails is 0. *Would you believe that? We need a prior!*
– Look at several independent trials. Consider N = 3 and m = 2 (N is number of trials, m is number of successes) and look at all ways to get 2 H and 1 T:
  * H H T $\rightarrow \mu\mu(1-\mu) = \mu^2(1-\mu)$
  * H T H $\rightarrow \mu(1-\mu)\mu = \mu^2(1-\mu)$
  * T H H $\rightarrow (1-\mu)\mu\mu = \mu^2(1-\mu)$
– $\begin{pmatrix} 3 \\ 2 \end{pmatrix} \mu^2(1-\mu) \rightarrow \begin{pmatrix} N \\ m \end{pmatrix} \mu^m (1-\mu)^{N-m} = \frac{N!}{(N-m)!m!} \mu^m (1-\mu)^{N-m}$
– This is the Binomial Distribution, gives the probability of m observations of $x = 1$ out of N independent trails
– So, what we saw is that we need a prior. We want to incorporate our prior belief. Let us place a prior on μ

$$Beta(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \tag{30}$$

$$E[\mu] = \frac{a}{a+b} \tag{31}$$

$$Var[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{32}$$

– Note: $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ and when x is an integer, then it simplifys to $x!$

– Calculation of the posterior, Take $N = m + l$ observations:

$$p(\mu|m,l,a,b) \propto Bin(m,l|\mu)Beta(\mu|a,b) \tag{33}$$
$$\propto \mu^m(1-\mu)^l \mu^{a-1}(1-\mu)^{b-1} \tag{34}$$
$$= \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{35}$$

– What does this look like? Beta: $a \leftarrow m + a$, $b \leftarrow l + b$
– So, whats to posterior?

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{36}$$

– **Conjugate Prior Relationship:** When the posterior is the same form as the prior
– Now we can maximize the (log of the) posterior:

$$\max_{\mu}(m+a-1)\ln\mu + (l+b-1)\ln(1-\mu) \tag{37}$$

$$\frac{\partial \mathscr{L}}{\partial \mu} = 0 = \frac{m+a-1}{\mu} - \frac{l+b-1}{1-\mu} \tag{38}$$
$$= (1-\mu)(m+a-1) - \mu(l+b-1) \tag{39}$$
$$= (m+a-1) - \mu(m+a-1) - \mu(l+b-1) \tag{40}$$
$$\mu = \frac{m+a-1}{m+a+l+b-2} \tag{41}$$

– This is the MAP solution. *So, what happens now when you flip one heads, two heads, etc.?*
– Discuss online updating of the prior. Eventually the data takes over the prior.