# CS 156a Final

Claire Goeckner-Wald

December 2, 2016

# Nonlinear Transforms

**1.** [e] None of the above.

The polynomial transform of order $Q = 10$ applied to $\chi$ of dimension $d = 2$ results in a $\mathbb{Z}$ space of dimensionality $\tilde{d} = 65$, not counting the constant coordinates $x_0 = 1$ and $z_0 = 1$. The polynomial transform of order $Q$ applied to some $x \in \chi = (x_1, x_2)$ consists of all variations of $x_1^i x_2^j$ with non-negative integers $i$ and $j$ satisfying $i + j \leq Q$. For example, the polynomial transform of order $Q = 3$ applied to $\chi$ of dimension $d = 2$ is

$$\begin{aligned} z &= (x_1^0 x_2^0, x_1^1 x_2^0, x_0^0 x_2^1, x_1^1 x_2^1, x_1^1 x_2^2, x_1^2 x_2^1, x_1^3 x_2^0, x_1^0 x_2^3); \\ &= (1, x_1, x_2, x_1 x_2, x_1 x_2^2, x_1^2 x_2, x_1^3, x_2^3). \end{aligned}$$

In this case, the dimensionality, $\bar{d}$, of the $\mathbb{Z}$-space is 7. (We do not count the constant coordinate $z_0$.) Then, for some polynomial transform of order $q$, we have $\tilde{d} = (\sum_{i=0}^{q} (q - i) + 1) - 1$. Therefore, for $Q = 10$, we have $\tilde{d} = (\sum_{i=0}^{10} (10 - i) + 1) - 1$, or $\tilde{d} = 65$.

# Bias and Variance

**2.** [d] $\mathcal{H}$ is the logistic regression model

**[a]** If $\mathcal{H}$ is a singleton, then the chosen hypothesis $g^{\mathcal{D}}$ will be the same, no matter the data set $\mathcal{D}$. Thus, when finding the average value of $g^{\mathcal{D}}$ with respect to $\mathcal{D}$, the average hypothesis must be the single hypothesis in $\mathcal{H}$. Therefore, $\bar{g} \in \mathcal{H}$.

**[b]** If $\mathcal{H}$ is the set of all constant, real-valued hypotheses, then the chosen hypothesis $g^{\mathcal{D}}$ for some data $\mathcal{D}$ will be a constant real number $b \in \mathbb{R}$. Therefore, when finding the average value of $g^{\mathcal{D}}$ with respect to $\mathcal{D}$, the expected value of $\bar{g}$ must also be some real-valued constant $\bar{b} \in \mathbb{R}$. Thus, since $\bar{b} \in \mathbb{R}$, then $\mathcal{H}$ must contain $\bar{b}$. Therefore, $\bar{g} \in \mathcal{H}$.

**[c]** If $\mathcal{H}$ is the linear regression model, then it represents hypotheses consisting of a polynomial (or, binomial) with real-valued coefficients. Therefore, when finding the average value of $g^{\mathcal{D}}$ with respect to $\mathcal{D}$, the expected value of $\bar{g}$ must also be some set of real-valued coefficients. Thus, since that set is real-valued, then $\mathcal{H}$ must contain it as a hypothesis. Therefore, $\bar{g} \in \mathcal{H}$.

**[d]** If $\mathcal{H}$ is the logistic regression model, then it represents hypotheses as sigmoidal functions. Then, when finding the chosen hypothesis $g^{\mathcal{D}}$ for some data $\mathcal{D}$, the hypothesis is a sigmoid ranging from 0 to 1. However, when finding the average value of $g^{\mathcal{D}}$ with respect to $\mathcal{D}$, the expected value of $\bar{g}$ is not necessarily a sigmoid. Given sigmoids $f(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ and $f(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$, then $h(x) = f(x) + g(x)$ is not necessarily a sigmoid. Since $\mathcal{H}$ contains sigmoids, $\bar{g}$ is not necessarily in $\mathcal{H}$.

# Overfitting

**3.** [d] *is false*

**[a]** *True*: Overfitting is the result of picking a hypothesis with a higher $E_{out}$ because of a misleadingly low $E_{in}$. Therefore, there must be two or more hypotheses with differing values of $E_{in}$, in order for us to be misled into overfitting.

**[b]** *True*: If there is overfitting, there must be two or more hypotheses that have different values of $E_{\text{out}}$ because overfitting results in choosing an incorrect hypothesis $g$ with a higher $E_{\text{out}}$ as a cause of a misleadingly lower $E_{\text{in}}$. Thus, there must be two or more hypotheses with different values of $E_{\text{out}}$ in order for overfitting to occur.

**[c]** *True*: If there is overfitting, there must be two or more hypotheses that have different values of $(E_{\text{out}} - E_{\text{in}})$. For example, consider the set $\mathcal{H}$ composed of hypotheses $h_1$ and $h_2$, each with out-sample errors $e_{\text{out},1}$ and $e_{\text{out},2}$ respectively, where $e_{\text{out},1} < e_{\text{out},2}$. Therefore, it is best if we choose that $g = h_1$. Overfitting occurs when $h_2$, the inferior hypothesis, has a misleading low value of $E_{\text{in}}$. Let each hypothesis have in-sample errors $e_{\text{in},1}$ and $e_{in,2}$. Then, in the case of overfitting, $e_{\text{in},1} > e_{\text{in},2}$. Thus, in the case of overfitting, we have $(e_{\text{out},1} - e_{\text{in},1}) < (e_{\text{out},2} - e_{\text{in},2})$. Therefore, if there is overfitting then two or more hypotheses that have different values of $(E_{\text{out}} - E_{\text{in}})$.

**[d]** *False*: We cannot always determine if there is overfitting using values of $(E_{\text{out}} - E_{\text{in}})$. Consider hypothesis set $\mathcal{H}$ with two hypotheses, $h_1$ and $h_2$. In this example, we have no overfitting because we select hypothesis $h_1$ for its superior $E_{in}$

$$e_{\text{in},1} = 0.2, e_{\text{out},1} = 0.5, (e_{\text{out},1} - e_{\text{in},1}) = 0.3 e_{\text{in},2} \qquad = 0.6, e_{\text{out},2} = 0.7, (e_{\text{out},2} - e_{\text{in},2}) = 0.1$$

Now, consider hypothesis set $\mathcal{H}'$ with two hypotheses, $h_1'$ and $h_2'$. In this example, assuming a naive learning algorithm, we have overfitting because we are misled by the misleading in-sample error of $h_1'$.

$$e_{\text{in},1} = 0.2, e_{\text{out},1} = 0.5, (e_{\text{out},1} - e_{\text{in},1}) = 0.3 e_{\text{in},2} \qquad = 0.3, e_{\text{out},2} = 0.4, (e_{\text{out},2} - e_{\text{in},2}) = 0.1$$

Since we are only given $(E_{\text{out}} - E_{\text{in}})$, we will similar values for both scenarios- however, one demonstrates overfitting, and the other does not. Thus, we cannot always determine the presence of overfitting given $(E_{\text{out}} - E_{\text{in}})$.

**[e]** *True*: We cannot determine overfitting based on one hypothesis only.

**4.** [d] *is true*

**[a]** *False*: Deterministic noise can occur with stochastic noise- the two are not existentially dependent.

**[b]** *False*: Deterministic noise depends on the hypothesis set, because we can construct certain hypothesis sets $\mathcal{H}$ that are prone to high deterministic noise. For example, if a hypothesis set $\mathcal{H}_2$, composed of exclusively second-order polynomials, attempts to fit a tenth-order polynomial target function with noise, then the deterministic noise will likely be high.

**[c]** *False*: Deterministic noise depends on the target function, as seen in the lower right-hand graph of the logo of this class. In this graph, the increasing target complexity results in a worse $E_{\text{out}}$, due to deterministic noise. For example, if we have a hypothesis set $\mathcal{H}_2$, composed of exclusively second-order polynomials, attempting to fit a variety of different $n$-order target functions, some target functions will result in a lower deterministic noise than others.

**[d]** *True*: Stochastic noise does not depend on the hypothesis set. Stochastic noise is a byproduct of the random flucuations of the given data. For example, stochastic noise may occur where human error is present in measurements performed to create the training data set.

**[e]** *False*: Stochastic noise does depend on the target function. Stochastic noise is random flucuations of given data set that is formed using the target function. Thus, if the target function changes, we may have different stochastic noise.

## Regularization

**5.** [a] $\boldsymbol{w}_{\text{reg}} = \boldsymbol{w}_{\text{lin}}$

In the case that the solution for $\boldsymbol{w}_{\text{lin}}$ is already constrained by $\boldsymbol{w}^\intercal\Gamma^\intercal\Gamma\boldsymbol{w} \leq C$, then we simply use that value of $\boldsymbol{w}_{\text{lin}}$ as our regularized weight, $\boldsymbol{w}_{\text{reg}}$. This is because $\boldsymbol{w}_{\text{lin}}$ is also found by minimizing $\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}^\intercal\boldsymbol{x}_n - y_n)^2$, so if $\boldsymbol{w}_{\text{lin}}$ satisfies the constraint as well, then $\boldsymbol{w}_{\text{reg}} = \boldsymbol{w}_{\text{lin}}$.

**6.** [b] translated into augmented error

**[a]** Writing a soft-order constraint as a hard-order constraint is not the point of the soft-order constraint.

**[b]** Soft-order constraints that regularize polynomial models can be translated into augmented error through a series of calculations.

$$\text{Minimizing } E_{\text{in}}(\boldsymbol{w}) = \frac{1}{N}(Z\boldsymbol{w} - \boldsymbol{y})^\intercal(Z\boldsymbol{w} - \boldsymbol{y}) \text{ subject to } \boldsymbol{w}^\intercal\boldsymbol{w} \leq C$$

$$\text{Can be solved by } \boldsymbol{w}_{\text{reg}} = (Z^\intercal Z + \lambda I)^{-1}Z^\intercal\boldsymbol{y}\text{, with } \lambda \text{ being the ``amount of regularization''}$$

**[c]** The value of the VC *dimension* does not determine the soft-order constraint. However, the VC *formulation* is equivalent to a soft-order constraint.

**[d]** Using soft-order constraints may result in an *increased* $E_{\text{in}}$ and decreased $E_{\text{out}}$ simultaneously. (This is kind of the goal of soft-order constraints.)

## Regularized Linear Regression

**7.** [d] 8 versus all
See attached code.

**8.** [b] 1 versus all
See attached code.

**9.** [e] The transform improves the out-of-sample performance of "5 versus all", but by less than 5%
See attached code.

**10.** [a] Overfitting occurs (from $\lambda = 1$ to $\lambda = 0.01$)
See attached code.

## Support Vector Machines

**11.** [e] 0, 1, -0.5

After plotting the points, the value of $z_2$ makes no difference in the classification. Therefore, the value $w_2$ is 0. If a point falls on the plane, it must be 0. So, we must find a $b$ and $z_1$ for this scenario.

$$\begin{aligned} z &= (0.5, 0) \text{ is on the desired line} \\ (1,0) \cdot (0.5, 0) + b &= 0 \\ 0.5 + b &= 0 \\ b &= -0.5 \end{aligned}$$

**12.** [c] 4-5
See attached code.

## Radial Basis Functions

**13.** [a] $\leq 5\%$ of the time
See attached code.

**14.** [a] or [e]
See attached code.

**15.** [a] or [e]
See attached code.

**16.** [b] or [d]
See attached code.

**17.** [b] or [c]

See attached code.

**18.** [a] or [e]

See attached code.

## Bayesian Priors

**19.** [b] The posterior increases linearly over [0,1]

We assume that $P(h = f)$ is uniform over $[0, 1]$. Thus, the probability that $h = f$ given no data is uniform. Given a single data point $y = 1$ (that is, a person had a heart attack), we know that $f! = 0$. But, we also know that it is still possible that $f! = 1$. Therefore, $P(f = 0|\mathcal{D}) = 0$ and $P(0 < f < 1|\mathcal{D})! = 0$. Thus, we can eliminate answer choices [a] and [d]. We now must determine whether $P(h = f|\mathcal{D})$ increases linearly or nonlinearly over [0,1].

$$P(h = f|\mathcal{D}) = \frac{P(\mathcal{D}|h = f)P(h = f)}{P(\mathcal{D})} \propto P(\mathcal{D}|h = f)P(h = f)$$

$P(h = f)$ is a uniform distribution over all hypotheses. $P(\mathcal{D}|h = f)$ is a linear distribution increasing over [0,1]. This is because, given that one person had a heart attack, the probability that you chose someone with a heart attack is equal to the probability that any given person has a heart attack, which is $h$, which equals $f$. Thus, as $h$ increases, so does $P(\mathcal{D}|h = f)$, linearly. Therefore, our posterior increases linearly over [0,1].

## Aggregation

**20.** [c] $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$ *is true*

**[a]** *False*: $E_{\text{out}}(g)$ can be worse than $E_{\text{out}}(g_1)$. Consider the case where $g_2$ is consistently farther from the target $y = f(\boldsymbol{x})$ for $\boldsymbol{x} \in \chi$ than $g_1$ is. Then, taking the average of outputs, $E_{\text{out}}(g)$, would be greater than $E_{\text{out}}(g_1)$.

**[b]** *False*: $E_{\text{out}}(g)$ can be worse than the minimum of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$. Consider the minimum to be that of $g_{\text{min}}$, which is either $g_1$ or $g_2$. The other shall be $g_{\text{max}}$. In this case, $g_{\text{min}}$, will, by definition, have a lower $E_{\text{out}}(g_{\text{min}})$ than that of $g_{\text{max}}$. Thus, $g_{\text{max}}$ tends to be farther from the target $y = f(\boldsymbol{x})$ for $\boldsymbol{x} \in \chi$ than $g_{\text{min}}$ is. Thus, $E_{\text{out}}(g)$ would be greater than $E_{\text{out}}(g_{\text{min}})$.

**[c]** *True*: $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$ because we are using mean *squared* error. This means that $E_{\text{out}}(g)$, even if higher than $E_{\text{out}}(g_1)$ or $E_{\text{out}}(g_2)$, cannot be worse than their average.

$$
\begin{aligned}
E_{\text{out}}(g) &= \mathbb{E}\big[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\big] = \mathbb{E}\big[(f(\boldsymbol{x}) - \frac{1}{2}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big))^2\big] \\
\frac{1}{2}(E_{\text{out}}(g_1) + E_{\text{out}}(g_2)) &= \frac{1}{2}(\mathbb{E}\big[(f(\boldsymbol{x}) - g_1(\boldsymbol{x}))^2\big] + \mathbb{E}\big[(f(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2\big]) \\
&= \frac{1}{2}\mathbb{E}\big[(f(\boldsymbol{x}) - g_1(\boldsymbol{x}))^2 + (f(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2\big] \\
&= \frac{1}{2}\mathbb{E}\big[\big(f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})g_1(\boldsymbol{x}) + g_1(\boldsymbol{x})^2\big) + \big(f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})g_2(\boldsymbol{x}) + g_2(\boldsymbol{x})^2\big)\big] \\
&= \frac{1}{2}\mathbb{E}\big[2f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big) + g_1(\boldsymbol{x})^2 + g_2(\boldsymbol{x})^2\big] \\
&= \mathbb{E}\big[f(\boldsymbol{x})^2 - f(\boldsymbol{x})\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big) + \frac{1}{2}g_1(\boldsymbol{x})^2 + \frac{1}{2}g_2(\boldsymbol{x})^2\big] \\
&= \mathbb{E}\big[f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})\frac{1}{2}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big) + \frac{1}{2}\big(g_1(\boldsymbol{x})^2 + g_2(\boldsymbol{x})^2\big)\big] \\
&= \mathbb{E}\big[f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})\frac{1}{2}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big) + \frac{1}{2}\big(g_1(\boldsymbol{x})^2 + 2g_1(\boldsymbol{x})g_2(\boldsymbol{x}) + g_2(\boldsymbol{x})^2\big) - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= \mathbb{E}\big[f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})\frac{1}{2}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big) + \frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 + \frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= \mathbb{E}\big[f(\boldsymbol{x})^2 - 2f(\boldsymbol{x})g(\boldsymbol{x}) + g(\boldsymbol{x})^2 + \frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= \mathbb{E}\big[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2 + \frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= \mathbb{E}\big[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\big] + \mathbb{E}\big[\frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= E_{\text{out}}(g) + \mathbb{E}\big[\frac{1}{4}\big(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\big)^2 - g_1(\boldsymbol{x})g_2(\boldsymbol{x})\big] \\
&= E_{\text{out}}(g) + \mathbb{E}\big[\frac{1}{4}(g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2\big]
\end{aligned}
$$

We use the linearity of expected values in the above calculation; $\mathbb{E}[u] + \mathbb{E}[v] = \mathbb{E}[u+v]$. Now, it suffices to show that $\mathbb{E}\big[\frac{1}{4}(g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2\big] \geq 0$.

$$
\begin{aligned}
\mathbb{E}\big[\frac{1}{4}(g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2\big] &\geq 0 \\
(g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}))^2 &\geq 0 \\
b = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) & \\
b^2 &\geq 0
\end{aligned}
$$

Thus, we have that $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.

**[d]** *False*: $E_{\text{out}}(g)$ does not have to fall between the inclusive interval $[E_{\text{out}}(g_1), E_{\text{out}}(g_2)]$. For example if $g_1$ is consistently overestimating the target $y = f(\boldsymbol{x})$ for $\boldsymbol{x} \in \chi$, and $g_1$ is consistently underestimating the target $y$, then taking the average by $g(\boldsymbol{x}) = \frac{1}{2}(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x}))$ may result in an $E_{\text{out}}(g)$ that is lower than that of both $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.